DYNAMICS OF EPISTASIS FROM DUPLICATE GENES TO GENOME-WIDE NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lin Xu

January 2012

DYNAMICS OF EPISTASIS FROM DUPLICATE GENES TO GENOME-WIDE NETWORKS

Lin Xu, Ph. D.

Cornell University 2012

Epistasis refers to the phenomenon that phenotypic consequences caused by mutation of one gene depend on one or more mutations at another gene. Epistasis is critical for understanding many genetic and evolutionary processes, including pathway organization, evolution of sexual reproduction, mutational load, ploidy, genomic complexity, speciation and the origin of life. However, the epistatic dynamics in biological systems under various internal and external perturbations are largely unknown. In this study, I firstly focused on exploring dynamics of epistasis between duplicate genes. I then investigated the properties of global epistatic networks under different traits. Finally I examined the dynamic changes of epistatic relations among genes under genetic and environmental perturbations.

I started my research by investigating the transcriptional dynamics of duplicate genes with negative epistasis under external perturbations. We found an interesting design principle that two epistatically interacting duplicate genes can acquire a fitness advantage under fluctuating environmental perturbations by achieving maximum expression levels asynchronously. Soon after finishing this project, instead of focusing on epistatic relations between duplicate genes, we analyzed a high-throughput experimental dataset investigating epistatic interactions among ~4,000 genes in baker's yeast, *Saccharomyces cerevisiae*. We showed that epistasis is prevalent (~13%

increase from the random expectations) and displays modular architecture among genes that underlie the same growth traits. More interestingly, our results indicate that hub genes responsible for the same growth traits tend to link epistatically with each other more frequently than random expectation.

When conducting these projects, we realized that few studies have examined the genome-wide dynamics of epistatic relations under different genetic and environmental perturbations, which might be due to limitations in screening epistatic relations for multiple mutants of the same genes or multiple environmental conditions under the current high-throughput experimental platforms. We addressed this issue theoretically by using Flux Balance Analysis (FBA), which involves the optimization of cellular objective functions and allows prediction of *in silico* flux values and/or growth. A series of unique properties of epistatic dynamics under various genetic and environmental perturbations have been revealed in our FBA simulations, and some of them are highly consistent with previous experimental studies.

BIOGRAPHICAL SKETCH

Lin Xu was born and raised in Jingmen, China. He graduated from Longquan High School in 2001 and began his undergraduate studies at Fudan University, Shanghai, China. In college he received a Jun-Zheng Undergraduate Research Award and Wangdao Scholarship and participated in a research group in the National Quantitative and Population Genetics Laboratory. Under the supervision of Dr. Zewei Luo, he investigated the genome-wide evolutionary pattern of gene length in eukaryotic and prokaryotic organisms.

Lin graduated with honors from Fudan University and moved to Ithaca, New York, in 2006, to attend graduate school at Cornell University to pursue a Ph. D. in Genetics and Development. Throughout the past five years, he has worked under the mentorship of Dr. Zhenglong Gu to explore the dynamics of epistatic networks. His research journey started with studies on negative epistasis between duplicate genes, and then moved to explore the genome-wide epistatic dynamics under genetic and environmental perturbations.

To my parents,

for your support and encouragement,

friendship, and never-ending love.

And to my wife,

for pushing me further than I thought possible,

for your tremendous patience,

for making me laugh every day,

and for your willingness to cook.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Zhenglong Gu, for his continual guidance. His teachings and efforts have helped me develop from a naive undergraduate to a knowledgeable scientific researcher. Zhenglong, for all you have done toward my scientific and personal development, I am exceedingly grateful and will never forget your help in my whole life. I would also like to thank my committee members, Drs. Andy Clark, Chris Myers and Haiyuan Yu, for their advice and support over the years, and Dr. Tim O'Brien for his help during the first two years of my life at Cornell.

My dear friends in the lab have made the past several years very joyful. I appreciate their unrelenting support. First, I would like to thank Mr. Brandon Baker, who is the smartest person I have ever met. His friendship greatly enriched my years in Ithaca. Working with him to explore interesting scientific questions is really an enjoyable journey. I appreciate a lot his help to my career and let us become best friend forever. Second, I want to thank Dr. Huifeng Jiang, who has really helped me a lot on many scientific issues, from exploring questions in various perspectives to how to write a scientific paper, and I am very happy that he has found an independent position to continue his own academic career. I also want to thank my lab mates, Dr. Zhe Wang and Mr. Kaixiong Ye for their helpful discussions on scientific projects and happy time we had together. It is really one of best labs I could imagine.

Finally, I want to thank my parents and my wife, Mrs. Hong Chen, for their love, support, and unwavering confidence in my abilities.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER ONE

Introduction

1. 1. Epistasis in biological systems

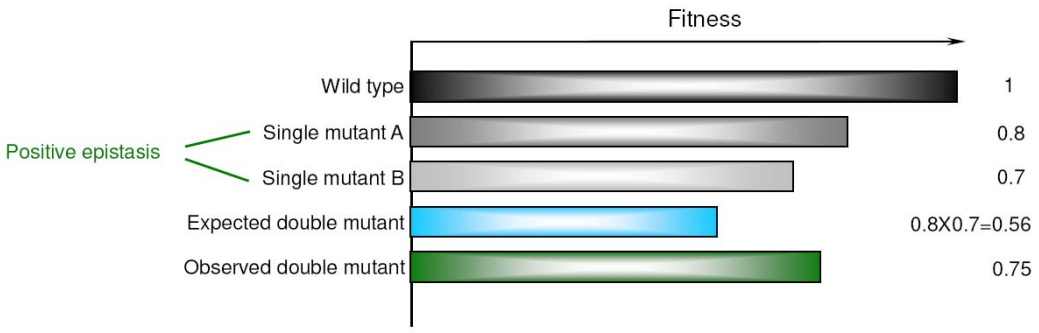
Epistasis between two deleterious mutations is positive when a double mutant causes a weaker mutational defect than predicted from individual deleterious mutations, and is negative when it causes a stronger defect (1, 2). As shown in Figure 1.1, positive and negative epistasis could be identified by comparing the expected and observed fitness values for the double mutants.

Epistasis has long been recognized as fundamentally important in understanding many genetic and evolutionary processes. For example, since epistasis is important in modifying the relationship between genotype and phenotype, illustrating epistasis underlying growth traits is essential for genetic mapping studies. Facilitated by the development of large-scale genotyping platforms, genome wide association studies (GWAS) have been widely used to identify candidate loci that might be causal for growth traits in various organisms, including human diseases. Although it is encouraging in identifying the risk loci underlying various complex diseases, GWAS is criticized by its low power in explaining a large proportion of heritability (3-5). Its utility in uncovering the genetic causes of complex diseases is questioned. Multiple reasons, one of which is epistasis among functionally related genes, were proposed to explain why the GWAS strategy can only explain a small amount of heritability in the experimental data (5). Therefore, illustrating the epistatic architecture underlying growth traits might lead to more efficient genetic mapping methods in identifying genetic factors that are responsible for phenotypic differences.

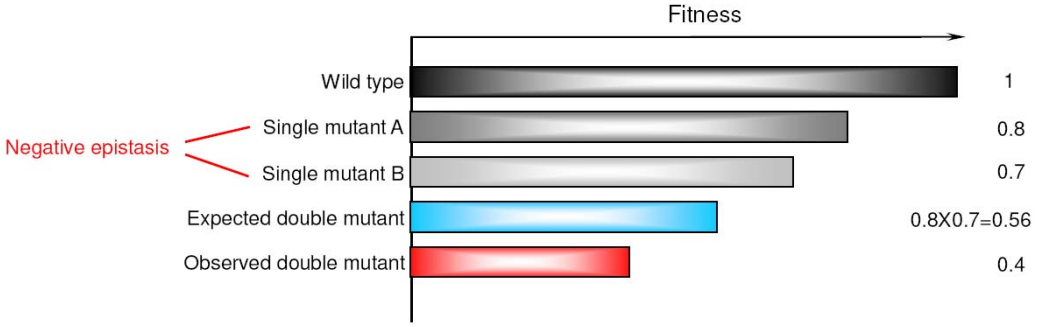
Besides the genetic mapping studies, it is also well established that epistasis is important for

Figure 1.1. A graphical representation of how epistatic interactions are inferred from a measurable single and double mutant fitness. (**A**) Positive epistatic interactions: Positive deviations from the value expected from the multiplicative model are scored as positive epistasis. In the figure, wild-type fitness is defined as 1. The fitness of two single mutants is 0.8 (single mutant A) and 0.7 (single mutant B), respectively. The expected fitness of the AB double mutant based on a multiplicative model would therefore be 0.56. The fitness of the observed AB double mutant is greater than expected (0.75); (**B**) Negative epistatic interactions: negative deviations from the value expected from the multiplicative model are scored as negative epistasis. The values for the single mutants are the same. However, the fitness of the observed AB double mutant is smaller than expected (0.4).

A


Fitness
Wild type
1
Positive epistasis
Single mutant A
0.8
Single mutant B
0.7
Expected double mutant
0.8X0.7=0.56
Observed double mutant
0.75

B


Fitness
Wild type
1
Negative epistasis
Single mutant A
0.8
Single mutant B
0.7
Expected double mutant
0.8X0.7=0.56
Observed double mutant
0.4

genomic complexity, mutational load, evolution of sex, speciation, and drug interactions (6-12). In addition, studies on epistasis at the genome level (epistatic interaction network, also called genetic interaction network) is getting more attention recently due to technological progress (reviewed in ref. 1 and 2). In the following sections, I am going to review the recent progress on studies of epistatic dynamics from duplicate genes to global epistatic interaction networks.

1. 2. Epistasis, genetic robustness and duplicate genes

One of the central tasks of genetics is to understand the origin and evolution of genetic robustness, and its relations to negative epistasis (4). Since duplicate genes have identical DNA sequences when they are produced and therefore have strong negative epistasis between each other, they are frequently used as an important resource to study the evolution of negative epistasis and genetic robustness (13-17). Recent progress in high-throughput approaches makes it possible to investigate the negative epistatic relations between duplicate genes by carrying out phenotypic analysis in null or conditional mutant alleles of duplicate genes on a genomic scale. Specifically, deletion libraries for model organisms like the budding yeast, *Saccharomyces cerevisiae*, and the RNAi-based screening in metazoans have greatly facilitated efforts to define epistatic relationship between duplicate genes (18-21).

The most surprising property revealed by systematical evaluation on the phenotypic consequences of gene loss is that most of gene disruptions seem not to be essential for viability (18-21). In yeast *S. cerevisiae*, ~80% of gene deletions have little or no detectable effect on

growth in rich media (18, 19). This number is ~90% in *E. coli* (22) and *C. elegans* (23). These facts highlight genetic robustness as one of the key properties in the biological system (24), which has been the subject of considerable interest (13, 19, 25-28).

In a pioneer work, Wagner (25) proposed two different mechanisms that are responsible for the genetic robustness: (a) genes have redundant paralogs, and mutations in one copy can be compensated by the other copy; (b) alternative pathways (non-duplicate genes), which function in a parallel fashion, could partially maintain the original functions when having malfunctions in one of them. Using genome-wide fitness assays of ~6,000 single deletion mutants in yeast *S. cerevisiae* (20), Gu et al (13) has shown that the existence of a paralog elsewhere in the genome significantly increases the chance that deletion of a given gene has little effect on growth in yeast *S. cerevisiae*, which provided the first genome-wide evidence to prove that the first mechanism is partially responsible for high proportion of dispensability for gene disruption. Following that, several independent genome-wide studies that compared fitness of double mutants for duplicates genes with that of single mutants further confirmed the conclusion in Gu et al (13) and supported the conclusion that the genome-wide excess dispensability among duplicates is partially due to backup compensation (14-17). The ability for duplicate genes to compensate each other's genetic mutations has also been shown in other species, such *C. elegans* (29, 30) and human (17). All these studies indicate that duplicate genes with negative epistasis play an important role in regulating global robustness of the biological systems.

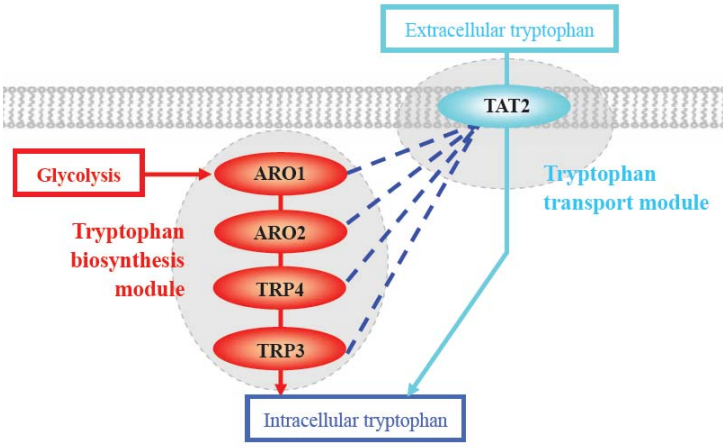
1. 3. Epistatic and physical interaction networks

Epistatic interaction network represents a network graph in which nodes are genes and edges are epistatic interactions, which was generated from a high-throughput experimental study (32). Complex epistatic interactions among more than 2 genes have not been experimentally measured in this study. Since epistatic interactions usually infer functional associations among genes (31, 32), it is expected that genes connected by epistatic interactions in such a network would confer relevant biological functions. Figure 1.2 displays two examples that we found in which epistatic interactions link different genes that execute specific biological functions. In the first case (Fig. 1.2A), the tryptophan permease TAT2, which is responsible for transporting extracellular amino acid tryptophan into cells, and the tryptophan biosynthesis module (ARO1, ARO2, TRP3, TRP4) could genetically compensate each other and therefore are closely connected by epistatic interactions. In the second case (Fig. 1.2B), protein modification genes UBP3 and BRE5 form a de-ubiquitination complex to cleave ubiquitin from many proteins to rescue biosynthesis (33). The SWR1 complex (SWC3, SWC5, SWR1, SWC2, ARP6) is known to be able to activate expression of many genes by remodeling chromatin structure (34). Genetic interactions closely connect these two functional modules. These individual cases prompted us to speculate whether genetic interaction networks have unique topologies that are important in maintaining organism genetic robustness.

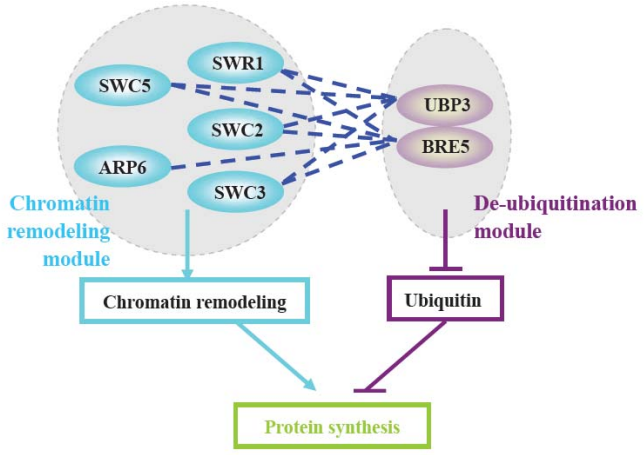
There are two major components in a cellular system: genetic information and genes that conduct and transfer the information. Correspondingly genes (or their encoded proteins) in the cellular

Figure 1.2. Individual cases for epistatic interactions that have real biological functions. Blue broken lines represent epistatic interactions. The lines terminated in arrowheads denote positive regulation, and the lines terminated in bars denote repression. Genes in the same shade are in the same functional module and genes with the same color function in the same pathways.

A


Extracellular tryptophan
TAT2
Glycolysis
ARO1
ARO2
TRP4
TRP3
Tryptophan biosynthesis module
Tryptophan transport module
Intracellular tryptophan

B


SWR1
SWC5
SWC2
ARP6
SWC3
UBP3
BRE5
Chromatin remodeling module
De-ubiquitination module
Chromatin remodeling
Ubiquitin
Protein synthesis

network are simultaneously linked by two kinds of interactions. On one hand, physical interaction networks, including the ones formed by protein-protein interaction, transcriptional regulatory interaction and metabolic interaction, dictate the architecture of a cell and play essential roles for executing genetic information. On the other hand, epistatic interactions focus on functional relationships between genes. Recent studies revealed that all physical interaction cellular networks share a unique topology in which the linkages among network hubs are systematically suppressed (35-38). How genes themselves are functionally associated in the cellular network, however, remains an unsolved issue. This is particularly intriguing in the light that disassortative topology can cause network to be fragile under perturbations (38-40), but biological systems are generally stable in face of heritable perturbations (41, 42).

Since the major biological role of physical interaction networks is to form concrete protein complexes, transcriptional circuits or metabolic pathway that are essential for biological signal transduction, disassortative topology that decrease the likelihood of crosstalk between different functional modules would provide a selective advantage to maintain the accuracy of biological signals (35). It was also proposed that the disassortative topology can localize effects of deleterious perturbation, which might increase genetic robustness. However, when we explored the organization principle of epistasis from a network perspective, we uncovered a unique architecture of preferential attachment to network hubs (Fig. 3.5), which is the opposite to the case of topologies of physical interaction networks. Moreover, as the architecture of physical interaction networks is important in preventing signal crosstalk, our results further indicate that

physical and genetic interaction networks could evolve unique – and in regard to gene relationships, even diametric – topologies to adapt to their biological roles and function in a cooperative manner.

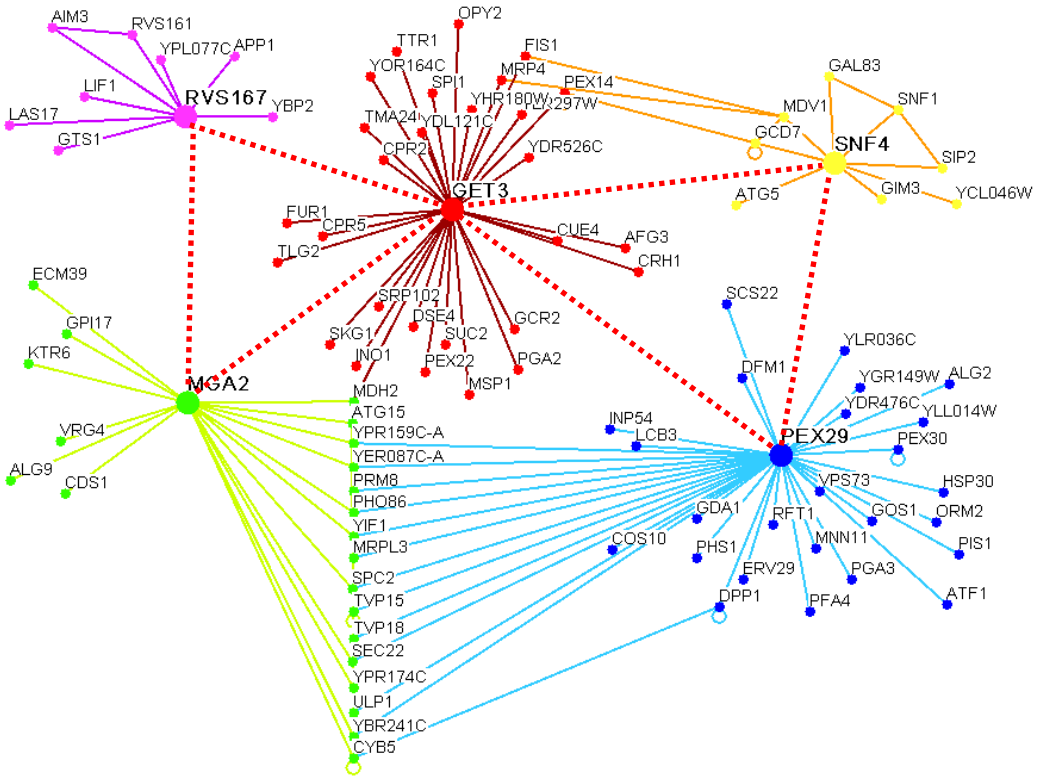
Figure 1.3 shows a concrete example in the real cellular network and elucidates the topological discrepancy between physical and genetic interaction networks. Five hub genes in both physical and epistatic interaction networks are involved in major functional modules that are responsible for stress response (GET3, MGA2 and RVS167) and peroxisome biogenesis (SNF4 and PEX29). Interestingly, a significant suppression of physical interactions among these genes (indeed no physical interactions) is observed. In contrast, these hub genes frequently link to each other by epistatic interactions, a feature that is consistent with the assortative topology in the epistatic interaction network.

1. 4. Epistatic dynamics under genetic and environmental perturbations

Previous experimental studies with a small number of genes have confirmed that the sign and magnitude of epistasis could vary under genetic and environmental perturbations (e.g. ref. 43). However, how epistatic interaction networks change under varying genetic and environmental perturbations are largely unknown.

Recently, considerable efforts have been put into the genome-wide measurements on the sign and magnitude of epistasis (31, 32, 44-58). A series of new high-throughput experimental

Figure 1.3. A concrete example of the topological discrepancy between physical (solid lines) and epistatic interactions (broken lines). Genes with the same color function in the same modules. For clarity, only epistatic interactions linking the hub genes are depicted.


AIM3
RVS161
YPL077C
APP1
LIF1
RVS167
YBP2
LAS17
GTS1
OPY2
TTR1
FIS1
YOR164C
SPI1
MRP4
PEX14
YHR180W
YLL297W
TMA24
YDL121C
CPR2
YDR526C
GET3
FUR1
CPR5
TLG2
CUE4
AFG3
CRH1
SRP102
DSE4
GCR2
SKG1
SUC2
INO1
PEX22
PGA2
MSP1
ECM39
GPI17
KTR6
MGA2
VRG4
ALG9
CDS1
MDH2
ATG15
YPR159C-A
YER087C-A
PRM8
PHO86
YIF1
MRPL3
SPC2
TVP15
TVP18
SEC22
YPR174C
ULP1
YBR241C
CYB5
GAL83
SNF1
MDV1
GCD7
SNF4
SIP2
ATG5
GIM3
YCL046W
SCS22
YLR036C
DFM1
YGR149W
ALG2
YDR476C
YLL014W
INP54
LCB3
PEX29
PEX30
VPS73
HSP30
GDA1
RFT1
GOS1
ORM2
COS10
PHS1
MNN11
PIS1
ERV29
PGA3
DPP1
PFA4
ATF1

platforms have been developed, such as synthetic genetic array (SGA) (31, 32), diploid-based synthetic lethality analysis with microarrays (dSLAM) (44), synthetic dosage-suppression and lethality screen (45-47) and epistatic miniarray profiles (EMAP) (48-50). In addition, genome-wide measurements on epistasis have been applied to various species (51-58).

A key issue for all these experiments is that the measurements of epistatic relations were mostly based on one mutant type for each gene and only under normal laboratory condition. Few studies constructed multiple mutants for single genes and examined the dynamics of epistatic relations among genes under multiple environmental perturbations. This situation is largely due to the lack of comprehensive collections of multiple mutants for the same genes and capability to conduct high-throughput experiments under a variety of distinct environmental conditions. More importantly, for genetic perturbations, even if the collections of a variety of genetic mutants are available, the enormous number of experiments for constructing double mutants and assaying fitness are still huge obstacles. The similar issue for examining epistatic dynamics under a large number of environmental perturbations also exists. As a consequence, the global landscape of epistasis under varying genetic and environmental perturbations remains largely uninvestigated.

1. 5. FBA predictions on global epistatic landscape

Given the importance of exploring global epistatic dynamics under genetic and environmental perturbations and the limitations for current experimental platforms, it is therefore necessary to develop the simulation technology for predicting and studying the global pattern of epistasis both

among various types of genetic mutants and under multiple distinct environmental perturbations.

Flux balance analysis (FBA) is a widely used approach for predicting growth rates of single and double mutants, as well as assessing epistatic relations among genes in the past decade (59-70). Based on a metabolic network that contain the known genes and reactions in an organism, FBA calculates the flow of metabolites through this metabolic network, thereby making it possible to predict the growth rate of an organism (71-73). In addition, FBA can be used to perform simulations under a variety of genetic and environmental perturbations by altering the constraints in a model, which provides a suitable mathematical framework for our purpose. The basic equations are shown as follows:

$$\boldsymbol{S}\boldsymbol{v} = \frac{d\boldsymbol{x}}{dt} = 0$$

$$\boldsymbol{v} = [v_1 v_2 \dots v_n]^T$$

$$\boldsymbol{v}_{\text{lb}} = [l_1 l_2 \dots l_n]^T$$

$$\boldsymbol{v}_{\text{ub}} = [u_1 u_2 \dots u_n]^T$$

$$\max_{\boldsymbol{v}} \boldsymbol{c}^T\boldsymbol{v} \quad \text{such that} \quad \boldsymbol{S}\boldsymbol{v} = 0 \text{ and } \boldsymbol{v}_{\text{lb}} \leq \boldsymbol{v} \leq \boldsymbol{v}_{\text{ub}}$$

S represents a numerical matrix, in which rows and columns correspond to metabolites and reactions in the reconstructed network respectively. **v** is the reaction flux with upper and lower bounds $\mathbf{v_{ub}}$ and $\mathbf{v_{lb}}$ respectively. Multiplying the stoichiometric matrix **S** by the flux vector **v**

equals to the concentration change over time ($d\mathbf{x}/dt$) (73). At steady state, the flux through each reaction is given by $\mathbf{Sv} = 0$, which defines a system of linear equations (73). FBA allows us to simulate a variety of genetic perturbations on an enzyme by constraining the fluxes corresponding to that enzyme. For example, if we want to simulate a deletion mutant of one enzyme, we can constrain its flux to be zero.

Conclusion: Epistasis represents an interesting research topic in the evolutionary field. I started my research by investigating the transcriptional dynamics of duplicate genes with negative epistasis under external perturbations. My research was later expanded to investigate the architecture of epistasis underlying specific traits. These experiences prompted me to further use FBA simulation to investigate how epistasis dynamically changes among genes under various genetic and environmental perturbations. Results in my researches have revealed several interesting properties for the distribution and dynamics of epistasis at the genome level, and call on further experimental effort to confirm these predictions. The evolutionary implications of my researches will also be discussed in each chapter.

References

1. Phillips PC (2008) Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9:855-867.

2. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet* 8:437-449.

3. Manolio TA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461:747-753.

4. McClellan J, King MC (2010) Genetic heterogeneity in human disease. *Cell* 141:210-217.

5. Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, Miletich JP, Ridker PM (2010) Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA* 303:631-637.

6. Presgraves DC (2007) Speciation genetics: epistasis, conflict and the origin of species. *Curr Biol.* 17:R125 – R127.

7. Hansen TF, Wagner GP (2001) Epistasis and the mutation load: a measurement theoretical approach. *Genetics* 158:477–485.

8. Kondrashov AS (1982) Selection against harmful mutations in large sexual and asexual populations. *Genet Res.* 40:325–332.

9. Musso G, et al. (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18:1092–1099.

10. Perez-Figueroa A, Caballero A, Garcia-Dorado A, Lopez-Fanjul C (2009) The action of purifying selection, mutation and drift on fitness epistatic systems. *Genetics* 1831:299 – 313.

11. Sanjuan R, Nebot MR (2008) A network model for the correlation between epistasis and genomic complexity. *PLoS One.* 37:e2663.

12. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I (2009) Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet.* 5:e1000578.

13. Gu Z, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63-66.

14. DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R (2008) Exposing the fitness contribution of duplicated genes. *Nat. Genet.* 40, 676-681.

15. Dean E J, Davis JC, Davis RW, Petrov DA (2008) Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLoS Genet.* 4(7), e1000113.

16. Musso G, Costanzo M, Huangfu M, Smith AM, Paw J, San Luis BJ, Boone C, Giaever G, Nislow C, Emili A, Zhang Z (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* Jul;18(7):1092-9.

17. Hsiao TL, Vitkup D (2008) Role of Duplicate Genes in Robustness against Deleterious Human Mutations. *PLoS Genet*. 4(3), e1000014.

18. Winzeler EA, et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901-906.

19. Giaever G, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387-391.

20. Steinmetz LM, et al (2002) Systematic screen for human disease genes in yeast. *Nat Genet* 31: 400–404.

21. Kamath RS, et al (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421: 231–237.

22. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2, 2006.0008.

23.Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* 408, 325-330.

24. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. *Science* 291, 1001-1004.

25. Wagner A (2000) Robustness against mutations in genetic networks of yeast. *Nat Genet* 24: 355–361.

26. Kafri R, Bar-Even A, Pilpel Y (2005) Transcription control reprogramming in genetic backup circuits. *Nat Genet* 37: 295–299.

27. Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays* 27: 176–188.

28. Papp B, Pal C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429: 661–664.

29. Gu X (2003) Evolution of duplicate genes versus genetic robustness against null mutations. *Trends in genetics* 19(7):354-6.

30. Vavouri T, Semple J, Lehner B (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends in Genetics* 24(10):485-8.

31. Tong AH, et al. (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808-813.

32. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327:425-431.

33. Cohen M, et al. (2003) Ubp3 requires a cofactor, Bre5, to specifically de-ubiquitinate the COPII protein, Sec23. *Nat. Cell. Biol.* 5:661-7.

34. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* 303:343-348.

35. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296: 910–913.

36. Petter Holme, Jing Zhao (2007) Exploring the assortativity-clustering space of a network's degree sequence. *PHYSICAL REVIEW E* 75:046111

37. Barabási AL, Oltvai ZN (2004) Network Biology: Understanding the Cells'Functional Organization. *Nature Reviews Genetics* 5:101-113

38. Newman MEJ (2002) Assortative mixing in networks. *Phys. Rev. Lett.* 89:208701.

39. Newman MEJ (2003) Mixing patterns in networks. *Phys. Rev. E* 67:026126

40. Vazquez A, Moreno Y (2003) *Phys. Rev. E* 67:015101(R)

41. Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle, J (2004) Robustness of cellular function.

Cell 118:675-685.

42. Wagner A (2005) Robustness and Evolvability in Living Systems. Princeton University Press, Princeton, NJ.

43. Remold SK, Lenski RE (2004) Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. Nature Genetics 36:423-426

44. Pan X, et al. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 124:1069-1081.

45. Measday V, Hieter P (2002) Synthetic dosage lethality. *Methods Enzymol* 350:316-326.

46. Measday V, et al. (2005) Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci USA* 102:13956-13961.

47. Sopko R, et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21:319-330.

48. Collins SR, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446:806-810.

49. Kornmann B, et al. (2009) An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science* 325:477-481.

50. Fiedler D, et al. (2009) Functional Organization of the *S. cerevisiae* Phosphorylation Network. *Cell* 136:952-963.

51. Sanjuán R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc Natl Acad Sci USA* 101:15376-15379.

52. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in HIV-1. *Science* 306:1547-1550.

53. Parera M, Ferna`ndez G, Clotet B, Martı´nez MA (2007) HIV protease catalytic efficiency effects caused by random single amino acid substitutions. *Mol Biol Evol* 24:382-387.

54. Elena SF, Lenski RE (1997) Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390:395-398.

55. Butland GM, et al. (2008) eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods* 5:789-795.

56. Typas A, et al. (2008) High-throughput, quantitative analyses of genetic interactions in *E. coli*. *Nat Methods* 5:781-787.

57. Roguev A, et al. (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 322:405-410.

58. deVisser JA, et al. (1997) Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. *Evolution* 51:1499-1505.

59. Harrison R, Papp B, Pal C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci USA* 104:2307-2312.

60. Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* 38:993-998.

61. Segrè D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* 37:77-83.

62. He X, Qian W, Wang Z, Li Y, Zhang J (2010) Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat Genet* 42:272-276.

63. Edwards JS, Ibarra RU, Palsson BØ (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19:125-130.

64. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 99:15112-15117.

65. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 102:7695-7700.

66. AbuOun M, et al. (2009) Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J Biol Chem* 284:29480-29488.

67. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* 33:164-190.

68. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic

reconstructions using *Escherichia coli*. *Nat Biotechnol* 26:659-667.

69. Fong SS, et al. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 91:643-648.

70. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res* 15:1365-1372.

71. Mo ML, Palsson BØ, Herrgard MJ (2009) Connecting extracellular metabolomic profiles to intracellular metabolic states in yeast. *BMC Syst Biol* 3:37.

72. Becker SA, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2:727-738.

73. Smallbone K, Simeonidis E (2009) Flux balance analysis: a geometric perspective. *J Theor Biol* 258:311-315.

CHAPTER TWO

Negative Epistasis between Asynchronously Regulated Duplicate Genes as a Strategy for Survival in Fluctuating Environments

2. 1. Summary

Organisms evolve in natural environments replete with changing stimuli. Understanding how organisms adapt to fluctuating environments is one of the primary goals in biology. Previous work on this topic focused on addressing the importance of genetic heterogeneity at the population level. How genetic regulatory circuits are designed by natural selection to cope with fluctuating environments is largely uncharacterized. By studying regulatory dynamics of the glucose transport system in the baker's yeast, *Saccharomyces cerevisiae*, we demonstrate an interesting design principle that two functionally interchangeable duplicate genes can acquire a fitness advantage under fluctuating environmental perturbations by achieving maximum expression level asynchronously. However, this advantage could come at a cost of competitiveness for organisms in constant environments. We were also able to show that most redundant duplicate genes are asynchronously regulated if they directly respond to external environments, which is not the case if the redundant duplicate genes function endogenously. To our knowledge, this is the first study to show a genome-wide strategy for genetic circuit regulation that could improve organismal fitness in fluctuating environments.

2. 2. Introduction

One of the central tasks for evolutionary biologists is to understand how organisms cope with fluctuating environments during evolution (1). Research on this issue can be linked to a study more than 40 years ago by Cohen who showed that seeds of an annual plant can either germinate or remain dormant in soil according to environmental conditions (2). This diverse field encompasses hundreds of articles spanning almost every branch of evolutionary biology. To name a few, fluctuating environments in a pre-biotic era were argued to be essential for formation of biotic materials (3, 4, 5). *K*- and *r*-selections were demonstrated to be different strategies for organisms to cope with fluctuating environments (6, 7). The fluctuating environments can also have great impact on important genetic features such as sex determination (8, 9, 10, 11), recombination (9, 12, 13), mutation rate (14, 15), and epistasis (16), etc. Recently, genetic regulatory noise was also proposed as an advantageous strategy for organisms in fluctuating environments (17, 18, 19).

Most, if not all, of these earlier works focused on studying how genetic heterogeneity at the population level enables adaptation in fluctuating environments. With different phenotypes of individuals in a population, organisms do not need an expensive system to be able to sense environmental fluctuation. Instead, cells could ‘blindly’ anticipate and survive environmental changes by having multiple phenotypes, each fit to a particular environment. This bet-hedging strategy might work, but it is very costly to the whole population because most individuals are doomed upon environmental switches. And clearly fitness of an individual would be increased if

its phenotype could physiologically adapt to environmental fluctuation. In this study we attempted to demonstrate how genetic regulatory circuits with redundant duplicate genes could enable organisms to adapt to fluctuating environments.

As a basic concept in molecular biology, genetic regulatory circuits describe "the gene and gene products that are involved in response to a signal" (20). In 1961, Jacob and Monod proposed for the first time the operon model for a genetic regulatory circuit, and introduced concepts like regulator gene and transcriptional repression (21, 22). Studying design principles for genetic regulatory circuits has become one of the central tasks in modern biology (reviewed in ref. 20). A well-established mathematical model to study transcriptional programs of a gene is:

$$\frac{dE_i}{dt} = A_i - k_i E_i \qquad (1)$$

where E_i is expression level of gene transcript i at time t; A_i is a transcription regulation term, which describes the activation effect by upstream signal(s), on transcript i at time t; and k_i is the decay rate for transcript i (23).

The mathematical models of the above formula (referred as Formula 1 in the following) are widely used to explore design principles of genetic regulatory circuits both in prokaryotic and in eukaryotic organisms. Shen-Orr et al. (24) applied this model to study design principles for feed-forward loops, single input modules (SIM) and dense overlapping regulons (DOR), which represent three basic genetic regulatory circuits in the transcriptional network in *E. coli*. In this

pioneering work, the authors successfully demonstrated that functional significance of genetic regulatory circuits could be understood by simulating their behaviors using circuit topologies and structures. The predictions based on these theoretical simulations were later confirmed by molecular experiments (25, 26). Similar studies were conducted on other genetic regulatory circuits, such as negatively autoregulated circuits (27), cascades (chains of regulatory reactions) (28), positive and negative feedback (29), flagellar systems (30) and various metabolic pathways (31).

In the present study, we applied similar modeling and simulation tools to study dynamic properties of genetic regulatory circuits that contain two redundant duplicate genes with asynchronous expression (*i.e.*, achieving maximum expression at different time upon stimuli), and proposed that such design could be effective for organisms to adapt to fluctuating environments. Analysis of transcriptional regulation of backup paralogs under their redundancy conditions reveals that most duplicate pairs that have redundant functions under exogenous conditions (by directly responding to external fluctuating environments) express asynchronously, whereas the ones with overlapping functions under endogenous conditions (unrelated to external environments) tend to be regulated synchronously. To investigate the possible selective advantage of asynchronous regulation of backup paralogs under fluctuating environments, we focused exclusively on two redundant glucose transporter genes. Our results suggest that asynchronous regulation of backup paralogs could provide a bet-hedging strategy for organisms to achieve proper expression states for survival when the environment fluctuates. We concluded

that both redundancy and asynchrony are necessary for such design in gene circuits that can successfully adapt to fluctuating environments.

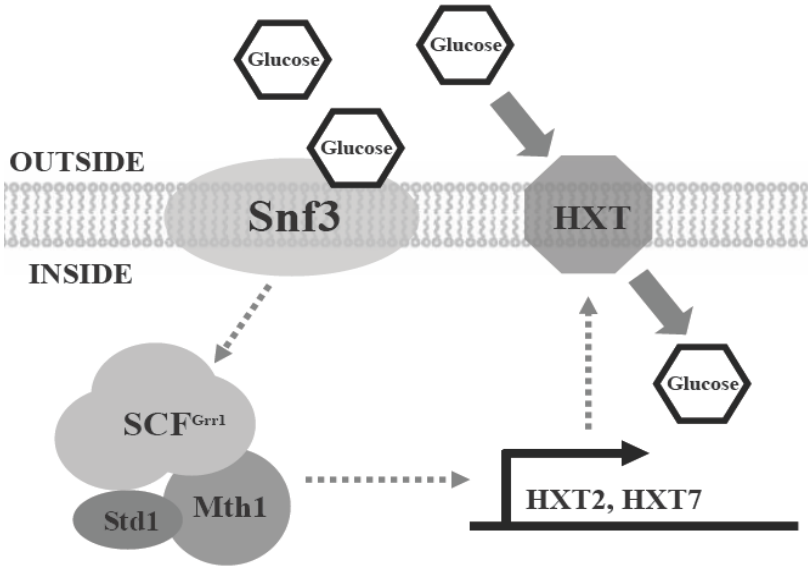
2. 3. Result

2. 3. 1. Transcriptional dynamics of redundant glucose transporter *HXT*2/7 genes

To investigate the possible selective advantage of asynchronous transcriptional regulation of backup paralogs under fluctuating environments, we focused on studying genes in the glucose transport system in *S. cerevisiae* due to their well-characterized functions. As shown in Figure 2.1, previous studies indicate that the *HXT*2 and *HXT*7 genes code for redundant high affinity glucose transporters (32-34). Because *HXT*2 and *HXT*7 are in the last step of glucose transporting system, expression levels of these genes could directly reflect amount of transported glucose. Transcriptional dynamics of these two genes under glucose pulses with different durations were monitored. To accomplish this, we first studied *HXT*2 and *HXT7* regulatory dynamics by real-time PCR under a short-term glucose pulse (~30 minutes). In this experiment, cells were transferred from 2g/L galactose medium to 0.2g/L glucose medium, glucose concentration (Figure 2.2A) and expression (Figure 2.2C) of *HXT*2 and *HXT*7 were monitored at different time points. The expression of *HXT*2 and *HXT*7 were then used to fit against a well established thermodynamic-kinetic principle for gene regulation (23).

As shown in Fig. 2.2C, expression level of both *HXT*2 and *HXT*7 genes increased rapidly within the first few minutes in glucose medium. However, when extracellular glucose drops below certain concentrations, expression levels of *HXT*2 and *HXT*7 starts decreasing in ~ 10 and ~15 minutes, respectively, which is much earlier than the time when total glucose was exhaustively used. The experimental results suggest that expression of *HXT*2 and *HXT*7 are well tuned by

Figure 2.1. Regulation of glucose sensing and transport. *Snf3p* senses low level glucose and generates a signal that activates SCF^{Grr1} complex. The SCF^{Grr1} complex mediates inactivation and degradation of *Mth1p* and *Std1p*, leading to expression of several *HXT* genes (33, 34). Among all *HXT* paralogs that respond to low glucose concentration, *HXT*2 and *HXT*7 express at high level and are responsible for the majority of glucose transport under this condition (34). Previous experiments showed that *HXT*2 and *HXT*7 have similar glucose transport efficiency and could compensate each other when one of them malfunctions (32, 33).


Glucose
Glucose
Glucose
OUTSIDE
Snf3
HXT
INSIDE
Glucose
SCF^Grr1
Std1
Mth1
HXT2, HXT7

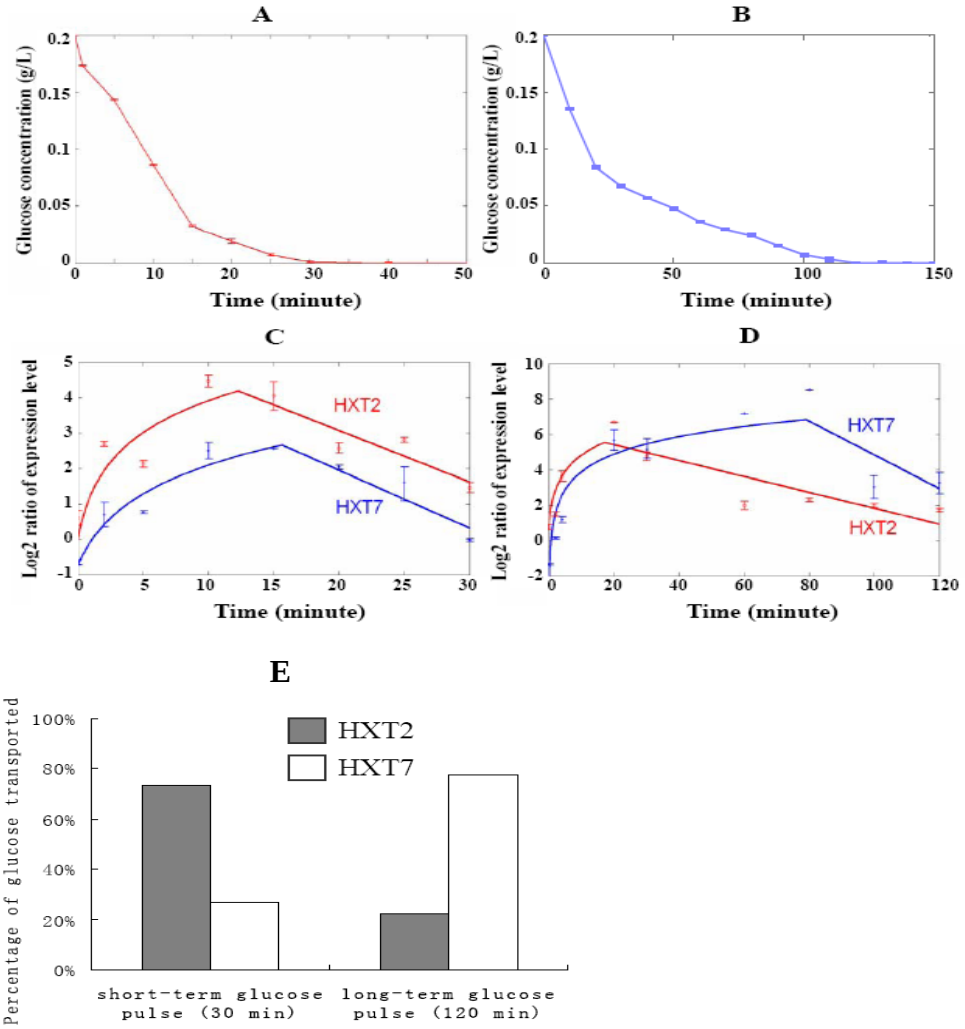
extracellular glucose concentrations. More importantly, our results indicate that the two genes have different transcriptional dynamics in this condition: the *HXT*2 gene achieved a higher total expression level than the *HXT*7 gene.

We then studied transcriptional profiles of *HXT*2 and *HXT*7 under long-term glucose pulse (~120 minutes) with ~10 fold decrease of initial cell concentration in the same glucose medium. Similar to short-term glucose pulse, expression levels of *HXT*2 and *HXT*7 started decreasing before the time when glucose was exhaustively used in the media (Figure 2.2B&2.2D). Interestingly, although *HXT*2 achieved the maximum expression level earlier than *HXT*7, the total amount of transcription for *HXT*7 was higher than that of *HXT*2 in this condition (Figure 2.2D). In summary, the *HXT*2 gene has a faster adaptation to glucose signal, but the *HXT*7 gene significantly increases its total expression if glucose availability lasts longer.

2. 3. 2. Asynchronous *HXT*2/7 regulation in fluctuating environments

Based on the obtained parameters for gene regulation dynamics, we predicted the total amount of glucose that can be transported by the *HXT*2 and *HXT*7 proteins under two glucose pulse conditions. As shown in Figure 2.2E, our results indicate that the *HXT*2 protein can transport more glucose in 30 minutes glucose pulse, thus a better choice for organism for glucose transport; when the glucose pulses is 120 minutes, however, the *HXT*7 protein outcompetes the *HXT*2 protein in glucose transporting.

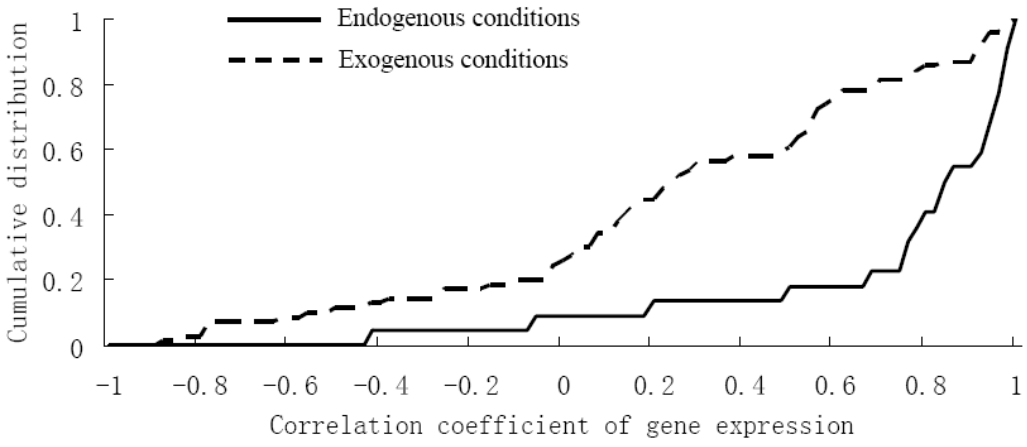
Figure 2.2. Glucose consumption and regulation of *HXT*2 and *HXT*7. Panels **A** and **B** show glucose consumption in short (30 minutes) and long (120 minutes) glucose pulses, respectively. Lines are produced by connecting different data points; **C** and **D** represent log2 ratio of expression levels for *HXT*2 (red line) and *HXT*7 (blue line) relative to the *ACT*1 gene in two conditions. The lines in C and D are produced by fitting expression data with a theoretical model, as described in Methods. **E** shows percentage of glucose transported by *HXT2* ($G_1/(G_1+G_2)$) and *HXT7* ($G_1/(G_1+G_2)$) under 30 and 120 minutes glucose pulses, whereas G_1 and G_2 are the amount of glucose transported by *HXT2* and *HXT7* proteins, respectively.


A
B
C
D
E
Glucose concentration (g/L)
Time (minute)
Log2 ratio of expression level
HXT2
HXT7
Percentage of glucose transported
short-term glucose pulse (30 min)
long-term glucose pulse (120 min)

2. 3. 3. Genome-wide regulatory pattern for redundant duplicate genes

It was previously shown that duplicate genes with redundant functions tend to have dissimilar gene expression (35-37). However, the global transcriptional program for backup paralogs under the conditions they have overlapping functions (termed as redundancy condition) remains unknown. To understand this issue, we automatically collected and manually checked references on redundant duplicate genes in *S. cerevisiae*. Available genome-wide microarray data for backup paralogs in their redundancy conditions were also recorded. ~90 pairs of backup paralogs, none of which are from high-throughput studies, were identified with experimental evidences and microarray gene expression data in their redundancy conditions. Correlation coefficients of mRNA gene expression for each pair of paralogs in their redundancy conditions were calculated. Interestingly, most duplicate genes whose redundancy conditions are unrelated to external environments (endogenous conditions, as defined in (38): cell cycle and sporulation) have closely correlated gene expression between paralogs, while gene pairs that directly respond to external environments (exogenous conditions, as defined in (38): oxidative stress, heat shock, high glucose, diauxic shift, NaCl tolerance and sobitol stress) likely show uncoupled expression (Figure 2.3). The discrepancy between two groups is highly significant (*Kolmogorov-Smirnov* test, $P<<0.0001$). This genome wide survey supports our conclusion that asynchronous (uncoupled) regulation of redundant duplicate genes is important for organism to properly respond to environmental stimuli.

Figure 2.3. Cumulative distributions for correlation coefficients of gene expression between backup paralogs under their redundancy conditions (endogenous v.s. exogenous). There are 6, 16, 9, 19, 16, 6, 14 and 5 redundant duplicate gene pairs for sporulation, cell cycle, oxidative stress, diauxic shift, heat shock, NaCl tolerance, high glucose and sorbitol stress, respectively. Redundant duplicate gene pairs whose redundancy conditions are exogenous show a significantly higher percentage of uncoupled transcriptional regulation than the gene pairs whose redundancy conditions are endogenous ($P<<0.0001$ by *Kolmogorov–Smirnov* test). The K_S values of duplicate gene pairs for the two categories do not show any differences (data not shown).


Endogenous conditions
Exogenous conditions
Cumulative distribution
1
0.8
0.6
0.4
0.2
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Correlation coefficient of gene expression

2. 4. Discussion

One of the central tasks for evolutionary biologists is to understand how organisms cope with fluctuating environments during evolution (1). Most, if not all, of the earlier work focused on studying how genetic heterogeneity at the population level enables adaptation in fluctuating environments (2, 7, 13, 14, 17-19). With different phenotypes of individuals in a population, organisms do not need an expensive system to be able to sense environmental fluctuation. Instead, cells could 'blindly' anticipate and survive environmental changes by having multiple phenotypes, each fit to a particular environment. This bet-hedging strategy might work, but it is very costly to the whole population because most individuals are doomed upon environmental switches. If an individual could physiologically adapt to environmental fluctuation, clearly the fitness of an individual would be increased. In this study we demonstrated how genetic regulatory circuits with redundant duplicate genes could contribute to this process.

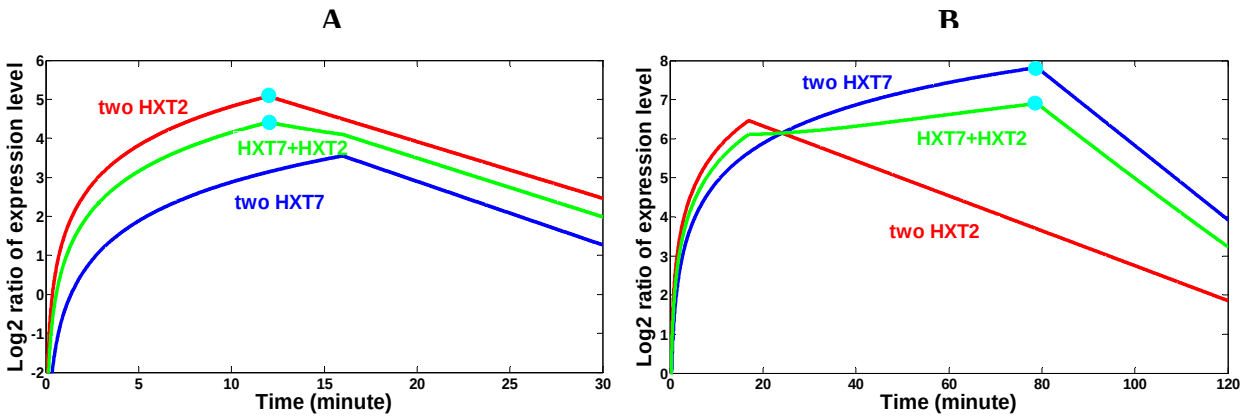
It is important to point out that advantage of regulating the *HXT*2 and *HXT*7 genes asynchronously in fluctuating environments is potentially associated with a cost. For example, strains containing asynchronously regulated *HXT*2 and *HXT*7 (wild type) have lower level of total glucose transporters than strains containing two synchronously regulated *HXT*2 genes under 30 minutes glucose pulse (Figure 2.4A). The same argument applies to the comparison between the wild-type and a postulated strain containing two *HXT*7 genes under 120 minutes glucose pulses (Figure 2.4B). Therefore, to increase chances of survival under fluctuating environments, wild-type strains might be suboptimal under specific constant conditions. This sub-optimality in

certain controlled conditions was also discovered in other biological processes, such as metabolic flux regulation (39).

It is interesting to explore how *HXT*2 and *HXT*7 could achieve their maximum expression level at different times. Several possible mechanisms, such as negative feedback to repress *HXT*7, positive feedback to amplify *HXT*2 (20), or transcription control reprogramming (35), could cause *HXT*7 to take longer than *HXT2* to achieve maximum gene expression. Experimental evidence is absent for supporting any one of the above models in the glucose transport system.

We assumed that transported glucose has a direct correlation with organism fitness. As glucose is the most important nutrient for organism growth, this assumption should be robust. Although our experiments were conducted in two conditions, it is reasonable to generalize our results that *HXT*2 and *HXT*7 could be optimal for short-term or long-term glucose pulse conditions, respectively. As the nutrient availability keeps fluctuating in environments, by asynchronously regulating *HXT*2 and *HXT*7, organisms might be able to cope with an uncertain nutrient level by hedging their bets, and thereby generate a much broader distribution of expression of *HXT* genes than synchronously regulated *HXT* genes. By using this tactic, the organism holds the promise that it could have larger odds in achieving proper expression state to survive when the environment fluctuates. To achieve this aim, both redundancy and asynchronous regulation for *HXT*2 and *HXT*7 are necessary and would be preserved. Interestingly, this gene-level strategy for adaptation to fluctuating environments resonate the bet-hedging strategy proposed at

Figure 2.4. Selective advantage and disadvantage of combing asynchronously regulated *HXT*2 and *HXT*7 in comparison to two synchronously regulated *HXT* genes (two *HXT*2 or two *HXT*7) in 30 minutes (**A**) and 120 minutes (**B**) of glucose pulses. The light blue point represents the state that has an optimal copy (the one that can transport more glucose) under that condition.


A
B
two HXT2
HXT7+HXT2
two HXT7
Log2 ratio of expression level
Time (minute)
two HXT7
HXT7+HXT2
two HXT2
Log2 ratio of expression level
Time (minute)

population-level, with which a population could take advantage of population heterogeneity among individuals in dealing with fluctuating environments (17, 19).

Our results indicate that genetic regulatory circuits with redundant duplicate genes could be important for organisms to adapt to fluctuating environments. In this scenario, natural selection only manipulates temporal regulation of existing genetic regulatory circuits without changing their original functional roles. By achieving regulatory divergence between duplicate genes but still keeping their gene functions, this hedging-bet strategy would be much safer than a strategy of only maintaining one expression state that is optimal to specific environments. Our unpublished results show that this design principle also applies to other redundant biological systems, such as redundant non-duplicate genes, pathways and network modules. By exploring properties of a biological system that enables adaptation to fluctuating environments, this study will advance our current understanding on evolution of complexity and how genetic robustness can contribute to organism adaptation to fluctuating environments (*i.e.* environmental robustness) in nature.

2. 5. Experimental Procedures

Strain and sample collection: Strain BY4741 (*Mat*a *his3*Δ *leu2*Δ *met15*Δ *ura3*Δ) was used in this study. Yeast cells were grown in minimal media with 2 g/liter galactose, 400 *rpm* at 30°C. Once the cultures achieved mid-log phase, cells were collected by centrifuge, washed three times to get rid of galactose, and diluted in minimum media without glucose. A glucose pulse that brought initial transient glucose concentration to 0.2 g/liter was injected into the flask. The initial cell concentrations were maintained at 10 OD/ml and 1 OD/ml to produce short-term (30 minutes) and long-term (120 minutes) glucose pulses, respectively. Samples for measuring glucose concentration were collected every 5 and 10 minutes for short and long term glucose pulse, respectively. Samples for measuring mRNA levels were collected at t =0, 2, 5, 10, 15, 20, 25, 30 minutes and at t=0, 2, 4, 20, 30, 60, 80, 100, 120 minutes for short and long-term glucose pulses, respectively.

Measurement of glucose concentration: A 1 ml sample was centrifuged at high speed to remove cells. The supernatant was stored at -20°C for further experiments. The residual glucose concentration in the growth medium was assayed by enzyme-coupled NADH oxidation reactions (Fisher Scientific. Cat. #. 9599191).

Measurement of gene transcription: Total cellular RNA was isolated using the hot phenol method as previously described (40) and treated with DNase I. The cDNA was produced with Qiagen Omniscript RT Kit (Cat. No. 205111). Quantitative PCR using SYBR green was

performed to measure transcription levels (primers for

*HXT*2: 2124U19:5'*TTGGTCGTCGTAAGTGTCT*3',

2261L22:5'*TGAATAAACAGGTAAAGACAAT*3'; and

*HXT*7: 1498U20:5'*CTACGGTTACGTTTTCATGG*3',

1665L20:5'*AATGGCTTGTCATCGTGAGT*3').

*ACT*1 gene was used as a control. Duplicate measurements from a single biological sample were conducted.

Kinetic modeling: As a basic concept in molecular biology, genetic regulatory circuits describe "the gene and gene products that are involved in response to a signal" (20). A well-established mathematical model to study transcriptional programs of a gene is:

$$\frac{dE_i}{dt} = A_i - k_i E_i \qquad (1)$$

where E_i is expression level of gene transcript i at time t; A_i is a transcription regulation term, which describes the activation effect by upstream signal(s), on transcript i at time t; and k_i is the decay rate for transcript i.

We built our model based on the Formula 1. As in (23), for a specific gene, the formal solution of the Formula 1 is

$$E(t) = \begin{cases} 0 & t \leqslant t_1 \\ (A/k_1)(1-e^{-k1(t-t1)}) & t_1 < t \leqslant t_2 \\ (A/k_2)(1-e^{-k2(t2-t1)})e^{-k2(t-t2)} & t_2 < t \end{cases} \qquad (2)$$

t_1 and t_2 represent activation and deactivation time of transcriptional regulation for the gene, respectively. A is a transcription regulation term, which describes the activation effect by upstream signal(s) and is assumed to be constant in the simulations (23, 24). k is the degradation rate for the transcript, which is assumed to be two constants in the activation (k_1) and deactivation processes (k_2) (23, 41, 42). To estimate these parameters, as in (4), we fit the actual mRNA expression level (C) from our experimental data (Figure 2.2) with the above solution by minimizing the sum of squares $\sum[E-C]^2$ in the space of parameters t_1, t_2, A and k with Matlab (Mathworks).

The total amount of glucose transported by *HXT* proteins (G_1 for *HXT*2 and G_2 for *HXT*7) was calculated by

$$G_i = \alpha \int_{t_1}^{D} r_t E_i \qquad (3)$$

Where α is the transporting rate of glucose for *HXT* proteins, which was estimated from (3), and r_t is the rate of translation per *HXT* mRNA molecule (corresponding to ribosome density) (43). Because the ratio of transcript length of *HXT*2 and is 0.95 (1,623 bp for *HXT*2 and 1,710 bp for *HXT*7), and ribosome density is determined by transcript length (44-46), it is reasonable to assume that *HXT*2 and *HXT*7 have nearly the same translational rate. In addition, because the decay time for the *HXT*2 and *HXT*7 protein molecules are ~600 and ~162 minutes, respectively (23), which are much longer than the time used in our experiments, for simplicity we assumed *HXT*2 and *HXT*7 proteins are stable in our simulations, a similar approach used in previous studies (47, 48). Based on these assumptions, we estimated relative amount of glucose

transported under short term (30 minute) and long term (120 minute) glucose pulses.

Genomic analysis: Duplicate genes are defined to have E value $<10^{-20}$ by BLASTP with standard parameters based on all-against-all search for verified ORFs in Saccharomyces Genome Database (SGD). In addition, the ratio of protein length between two paralogs should be smaller than 1.33 (35). Perl scripts were developed (available upon request) to collect references in PubMed that are related to *S. cerevisiae* duplicate genes with the key words that indicate redundant functions, such as "*redundant*", "*isoform*", "*overlapping*" etc. We then manually checked each reference to identify experimental evidences with the following criteria: (1) Two duplicate genes share the same upstream signal(s) and downstream target(s); (2) Clear molecular experimental evidences, none of which are from high-throughput studies, show that strains containing double knockouts display more severe defect than expected from two single knockouts, or these two duplicate genes function interchangeably under certain conditions. The experimental conditions (redundancy condition) under which each duplicate pair meets the above criteria were recorded. It is important to point out that these duplicate genes do not have 100% overlapping functions. Instead, they only show overlapping functions in the redundancy conditions.

Due to availability of the microarray data, duplicate genes whose redundancy conditions match one of the following microarray experiments were used in our analysis (endogenous conditions: cell cycle (49), sporulation (50); exogenous conditions: oxidative stress (51), heat shock (51), high glucose (52), diauxic shift (51), NaCl tolerance (51) and sorbitol stress (51). Endogenous

and exogenous conditions were defined as in (38).

References

1. Lewontin RC (1973) Population genetics. *Ann Rev Genetics* 7:1-18.

2. Cohen D (1966) Optimizing reproduction in a randomly varying environment. *J Theor Biol* 12:119-129.

3. Lahav N (1991) Prebiotic co-evolution of self-replication and translation or RNA world? *J Theor Biol* 151:531-539.

4. Lahav N, Nir S, Elitzur AC (2001) The emergence of life on Earth. *Prog Biophys Mol Biol* 75:75-120.

5. Lahav N, White D, Chang S (1978) Peptide formation in the prebiotic era: thermal condensation of glycine in fluctuating clay environments. *Science* 201:67-69.

6. Stearns SC (1976) Life-history tactics: a review of the ideas. *Q Rev Biol* 51:3-47.

7. Tuljapurkar S (1990) Delayed reproduction and fitness in variable environments. *Proc Natl Acad Sci U S A* 87:1139-1143.

8. Dooren TJ, Leimar O (2003) The evolution of environmental and genetic sex determination in fluctuating environments. *Evolution* 57:2667-2677.

9. Gandon S, Otto SP (2007) The evolution of sex and recombination in response to abiotic or coevolutionary fluctuations in epistasis. *Genetics* 175:1835-1853.

10. Leimar O, Dooren TJ, Hammerstein P (2004) Adaptation and constraint in the evolution of environmental sex determination. *J Theor Biol* 227:561-570.

11. Robson AJ, Bergstrom CT, Pritchard JK (1999) Risky business: sexual and asexual reproduction in variable environments. *J Theor Biol* 197:541-556.

12. Bourguet D, Gair J, Mattice M, Whitlock MC (2003) Genetic recombination and adaptation to fluctuating environments: selection for geotaxis in *Drosophila melanogaster*. *Heredity* 91:78-84.

13. Charlesworth B (1976) Recombination modification in a flucturating environment. *Genetics* 83:181-195.

14. Gillespie JH (1993) Substitution processes in molecular evolution I Uniform and clustered substitutions in a haploid model. *Genetics* 134:971-981.

15. Lamb BC, Mandaokar S, Bahsoun B, Grishkan I, Nevo E (2008) Differences in spontaneous mutation frequencies as a function of environmental stress in soil fungi at "Evolution Canyon," Israel. *Proc Natl Acad Sci U S A* 105:5792-5796.

16. Gillespie JH, Langley C (1976) Multilocus behavior in random environments I Random Levene models. *Genetics* 82:123-137.

17. Acar MJ, Mettetal T, Oudenaarden VA (2008) Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet* 40: 471-475.

18. Kitano H (2002) Systems biology: A Brief Overview. *Science* 295:1662-1664.

19. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. *Science* 309: 2075-2078.

20. Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet* 5: 34-42.

21. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318-356.

22. Jacob F, Monod J (1961) On the regulation of gene activity. *Cold Spring Harb Symp Quant Biol* 26:193-211.

23. Sásik R, Iranfar N, Hwa T, Loomis WF (2002) Extracting transcriptional events from temporal gene expression patterns during Dictyostelium development. *Bioinformatics* 18:61-66.

24. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 31:64-68.

25. Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *J Mol Biol* 356:1073-1081.

26. Mangan S, Zaslaver A, Alon U (2002) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* 334:197-204.

27. Rosenfeld, N, Elowitz, MB & Alon, U (2003) Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* 323:785-793.

28. Rosenfeld N, Alon U (2003) Response delays and the structure of transcription networks. *J Mol Biol* 329:645-648.

29. Süel G, Garcia-Ojalvo J, Liberman L, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* 440:545-550.

30. Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG, Alon U (2001) Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* 292:2080-2083.

31. Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, Alon U (2004) Just-in-time transcription program in metabolic pathways. *Nat Genet* 36:486-491.

32. Reifenberger E, Freidel K, Ciriacy M (1995) Identification of novel HXT genes in *Saccharomyces cerevisiae* reveals the impact of individual hexose transporters on glycolytic flux. *Mol. Microbiol.* 16:157-167.

33. Reifenberger E, Boles E, Ciriacy M (2004) Kinetic characterization of individual hexose transporters of *Saccharomyces cerevisiae* and their relation to the triggering mechanisms of glucose repression. *Euro. J. Biochemistry* 245:324-333.

34. Petit T, Diderich JA, Kruckeberg AL, Gancedo C, Dam KV (2000) Hexokinase regulates kinetics of glucose transport and expression of genes encoding hexose transporters in *Saccharomyces cerevisiae*. *J. Bacteriol.* 182:6815-6818.

35. Kafri R, Bar-Even A, Pilpel Y (2005) Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* 37:295-299.

36. Kafri R, Levy M, Pilpel Y (2006) The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci. USA* 103:11653-11658.

37. Kafri R, Springer M, Pilpel Y (2009) Genetic redundancy: new tricks for old genes. *Cell* 136:389-392.

38. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431:308-312.

39. Fischer E, Sauer U (2005) Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat. Genet.* 37:636-640.

40. Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C (2007) Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol.* 5:e90.

41. Tucker M, Parker R (2000) Mechanisms and control of mRNA decapping in *Saccharomyces cerevisiae. Annu. Rev. Biochem.* 69:571-595.

42. Wilusz CJ, Wormington M, Peltz SW (2001) The cap-to-tail guide to mRNA turnover. *Nat. Rev. Mol. Cell Biol.* 2:237-246.

43. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK (2006) Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. USA* 103:13004-13009.

44. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T (2004) Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell Proteomics* 3:1083–1092.

45. Preiss T, Baron-Benhamou J, Ansorge W, Hentze MW (2003) Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat. Struct. Biol.* 10:1039–1047.

46. MacKay VL, Li X, Flory MR, Turcott E, Law GL, Serikawa KA, Xu XL, Lee H, Goodlett DR, Aebersold R, et al. (2004) Gene expression in yeast responding to mating pheromone: Analysis by high-resolution translation state analysis and quantitative proteomics. *Mol. Cell Proteomics* 3:478–489.

47. Noe DA, Delenick JC (1989) Quantitative analysis of membrane and secretory protein processing and intracellular transport. *J. Cell Sci.* 92:449–459.

48. Wang X, Hao N, Dohlman HG, Elston TC (2006) Bistability, Stochasticity, and Oscillations in the Mitogen-Activated Protein Kinase Cascade. *Biophys. J.* 90:1961-1978.

49. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2:65–73.

50. Chu S, DeRisi J, Eisen M, MulhollJ, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.

51. Aach J, Rindone W, Church GM (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.* 10:431-45.

52. Ronen M, Botstein D (2006) Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proc. Natl. Acad. Sci. USA* 103:389-394.

CHAPTER THREE

Exploring Epistatic dynamics and Genetic Architecture of Growth Traits Revealed by Global Epistatic Interactions

3. 1. Summary

Epistasis has long been recognized as fundamentally important in understanding the structure, function and evolutionary dynamics of biological systems. Yet, little is known about how it is distributed with respect to specific traits. Based on a global map of epistatic interactions in baker's yeast, *Saccharomyces cerevisiae*, we show that epistasis is prevalent (~13% increase from random expectation) and displays modular architecture among genes that underlie the same growth traits. More interestingly, our results indicate that hub genes responsible for the same growth traits tend to link epistatically with each other more frequently than random expectation. Our results provide a genome-wide perspective on the genetic architecture of growth traits in a eukaryotic organism.

3. 2. Introduction

Complex traits that vary in populations of human and other organisms are determined by multiple genetic factors. An individual genetic factor might only contribute a modest amount to the total variation observed in a trait over the entire population (1-3). Genetic factors contributing to the same traits usually affect each other's phenotypic outcome, a phenomenon called epistasis (3, 4). How epistatic interactions among genetic factors are distributed underlying the same complex trait remains largely unknown (5). Here we use growth traits in yeast as models to study this issue.

It is also well established that epistasis is important for the evolution of sex (6-8), speciation (9), mutational load (10), ploidy (11, 12), genetic drift (13), genomic complexity (14), drug resistance (15), and human disease (5). In model organisms, illustrating epistatic interactions also enables dissection of functional relationship between genes (16-20). Understanding the distribution of epistasis underlying complex traits is therefore important for various fields.

Individual studies pointed out a prominent role for epistasis in genetic control of complex traits (21-24). However, a comprehensive understanding of epistasis underlying complex traits can only be achieved by reconstructing a global map of epistasis. Yeast provides a great model system to address this issue due to its abundant functional genomic data. Here we examined the distribution and prevalence of epistasis underlying growth traits of yeast in different conditions. We firstly identified genes which contribute to growth under each of 354 conditions (25). We

then extracted sub-networks of epistasis among the contributing genes in each of 354 conditions from the genome-wide epistatic network (26). Novel characteristics for the genetic architecture of growth traits are described. Although the epistasis used in this study was generated from yeast gene deletion mutants, and the complex traits used were measured from yeast growth in specific laboratory conditions, both of which might be different from the real scenario in nature, our results provide the first glimpse of the genome-wide organization of epistasis underlying complex traits. The implication of our results on gene pleiotropy is also discussed.

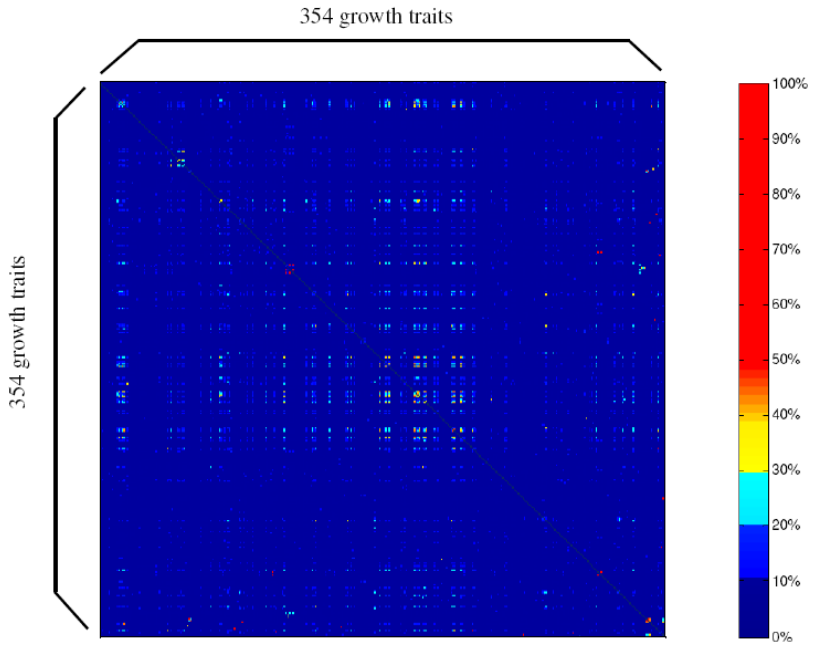
3. 3. Results and Discussions

3. 3. 1. Prevalent and modular epistasis

In order to study the genetic architecture of growth traits, we firstly identified genes that are responsible for growth traits. Based on a genome-wide screen for growth defects of ~6,000 *S. cerevisiae* gene deletion mutants in 354 distinct growth conditions, genes that contribute to growth in each condition were defined as those genes heterozygous deletion of which significantly affect organism growth in that condition (25). To ensure that these 354 conditions represent independent growth traits, we calculated the overlap of contributing genes between any two of the 354 conditions. As shown in Figure 3.1, 96% of the comparisons between any two conditions have less than 10% overlap of contributing genes, and 99% of the comparisons have less than 20% overlap, indicating that most of the 354 conditions are functionally independent. In addition, we took advantage of epistatic interaction data in yeast from a recent study (26), in which epistatic interactions are examined among more than five million gene pairs in *S. cerevisiae*. Sub-networks with epistatic interactions among contributing genes for each of 354 growth conditions were reconstructed.

While most people agree that epistasis plays an important role in the genetic architecture of complex traits, there is a disagreement about how common epistatic interactions are within genes that contribute to the same trait (5, 27). Using the reconstructed 354 epistasis sub-networks, we found that when two genes are responsible for the same growth trait, on average 3.6% of them are linked by an epistatic interaction. We then conducted a simulation by keeping the number of

Figure 3.1. The studied growth traits are functionally independent. The overlap of contributing genes between any two conditions was calculated. Color represents the percentage of contributing genes that overlap between two conditions (color scheme is shown to the right). The grey diagonal line represents comparison between each trait and itself, which is not counted in the figure. In total, 96% of the comparisons have less than 10% overlap of contributing genes, and 99% of the comparisons have less than 20% overlap of contributing genes.

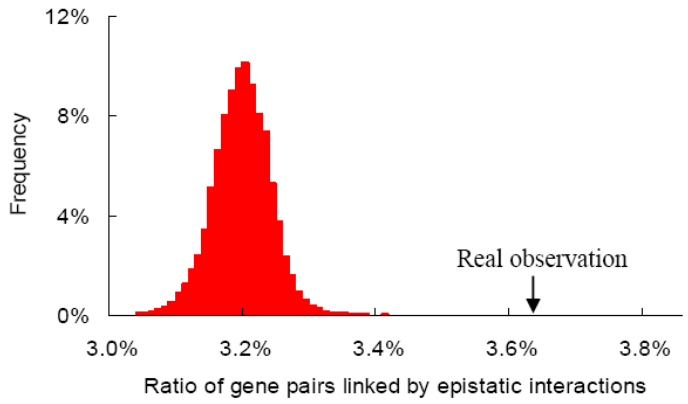

354 growth traits
354 growth traits
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

genes responsible for each trait as a constant, but randomly choosing genes to be responsible for each trait (repeated 100,000 times). In each iteration, we also calculated the fraction of gene pairs connected by epistatic interactions. As shown in Figure 3.2A, a significantly higher ratio was observed for the real experimental data than that of random expectation (~3.2%, Figure 3.2A, $P < 10^{-5}$), indicating that epistasis is enriched among genes responsible for the same biological traits. It is also noteworthy that the increase of epistasis (~13% more than random expectation) among contributing genes underlying the same growth trait in yeast are not dramatic, which may be due to the possibility that epistasis among genes vary in different conditions.

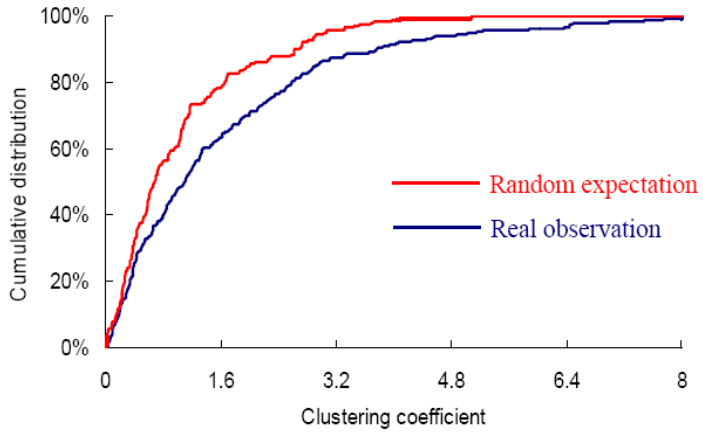
Previous studies proposed that gene pairs linked by epistatic interactions would be predictive of participation in common cellular functions (26, 28). However, although our above result is consistent with this expectation, it has never been shown before that the contributing genes underlying the same traits are also enriched with epistatic interactions. To further understand whether genes that contribute to the same growth traits are closely connected by epistasis, we calculated the average clustering coefficient, a network parameter that reflects the tightness of connection for a group of genes by immediate interactions (29), for genes that underlie each growth trait. The larger the clustering coefficient is, the more interlinked the group of genes are. For each of the 354 biological traits, we calculated the average clustering coefficient among its contributing genes. For comparison, we also calculated the average clustering coefficient for each trait in each of the above 100,000 simulations. Figure 3.2B shows the cumulative distributions of the clustering coefficients for real observation and random simulations. Our

Figure 3.2. Prevalent and modular epistasis in the genetic architecture of growth traits. (**A**) The distribution (red color) represents the average ratio of contributing gene pairs that are linked by epistatic interactions in 354 traits based on random simulations (repeated 100,000 times). The arrow indicates the average ratio of contributing gene pairs that are linked by epistatic interactions in 354 traits based on real experimental data. (**B**) The empirical cumulative distribution of the clustering coefficients for experimental observations (all 354 traits, blue curve) and random simulations (repeated 100,000 times for 354 traits, red curve). The *Kolmogorov–Smirnov* test indicates that the two distributions are significantly different ($P = 2x10^{-31}$).

A


12%
8%
4%
0%
Frequency
Real observation
3.0%
3.2%
3.4%
3.6%
3.8%
Ratio of gene pairs linked by epistatic interactions

B


100%
80%
60%
40%
20%
0%
Cumulative distribution
Random expectation
Real observation
0
1.6
3.2
4.8
6.4
8
Clustering coefficient

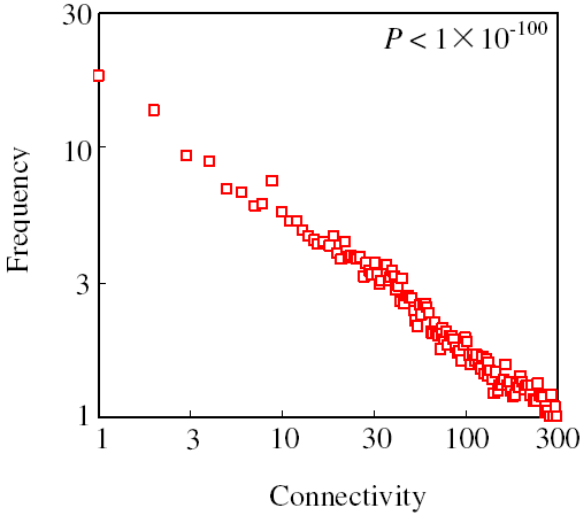
result indicates that genes underlying the same biological traits tend to be closely interconnected by epistatic interactions (*Kolmogorov-Smirnov test*, $P = 2x10^{-31}$).

3. 3. 2. Assortative characteristic of epistatic interactions for growth traits

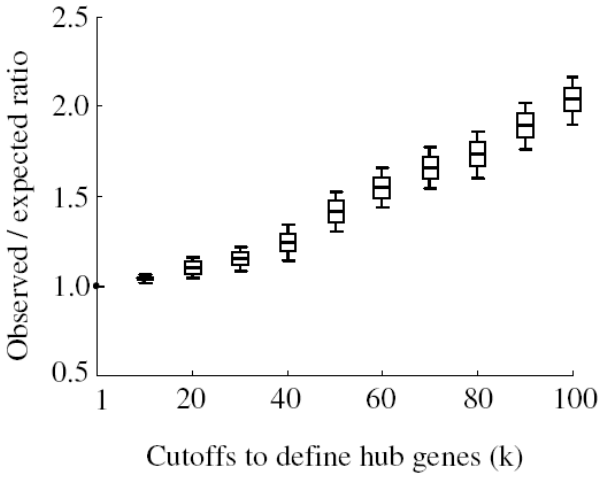
We further investigated how epistasis is distributed among the contributing genes for each growth trait. Many biological networks are scale-free, meaning that the degree (connectivity in the network) in these networks usually follows a so-called "power law" distribution (29). To examine whether the epistatic interactions among the contributing genes that underlie biological traits also display the scale-free characteristic, we calculated the connectivity for all contributing genes in each of the 354 sub-networks, respectively. We then investigated the distribution of degrees that were averaged over all 354 sub-networks. Figure 3.3A confirms that epistatic interactions underlying growth traits follow the power-law distribution. Contributing genes underlying most individual traits also show a similar pattern (Figure 3.4). How is epistasis distributed among the contributing genes with different connectivity? To answer this question, for each trait we first computed the number of epistatic interactions (N_{ko}) among the contributing genes that have more than k interactions in each of the observed epistasis sub-networks. Randomized versions of the epistasis sub-network were also generated for that trait, in which all the contributing genes have the same degree as the real epistasis sub-network, but the epistatic interactions between the contributing genes are randomly connected. For each trait, we calculated the average number of interactions (N_{ks}) among the contributing genes that have more than k interactions from 1,000 randomly generated networks. For each value k, we computed the

Figure 3.3. Assortative genetic architecture of growth traits. (**A**) The degree distribution of epistatic networks over 354 biological traits. MATLAB (Mathworks) was used to fit the regression and the small *p* value indicates that the network degree displays the scale-free characteristic. (**B**) Average ratio of observed/expected number of epistatic interactions among 354 traits. For each epistasis sub-network, the number of epistatic interactions among all contributing genes that have more than *k* epistatic interactions was calculated (observed numbers). The epistatic interaction in the sub-network was randomized and the average number of epistatic interactions among all contributing genes that have more than *k* epistatic interactions was also calculated from 1,000 random simulations (expected numbers). The bands, boxes and whiskers represent the means, ± 1 standard errors and ± 95% confidence intervals, respectively.

A


30
10
3
1
Frequency
$P < 1 \times 10^{-100}$
1
3
10
30
100
300
Connectivity

B


2.5
2.0
1.5
1.0
0.5
Observed / expected ratio
1
20
40
60
80
100
Cutoffs to define hub genes (k)

average ratio of N_{ko}/N_{ks} among all 354 traits. As depicted in Figure 3.3B, the ratio of N_{ko}/N_{ks} increases with k, indicating that epistasis is enriched among the contributing genes that are highly connected in the epistatic networks. When the ribosomal proteins and chaperones, which might represent universal hubs in the epistatic interaction network, are excluded, the pattern still holds (Figure 3.5).

When epistatic interactions among genes that affect organism growth in a particular condition are investigated, the contributing hub genes for this trait are by definition more likely than the non-hub genes to be linked by epistatic interactions. However, this increase in connectivity for the hub genes could be due to increased interactions either linking to other hub genes, or linking to non-hub genes. Indeed, the unique network architecture of enriched epistatic interactions among the hub genes revealed in this study, which is termed as "assortative" in social networks (30), is surprising because all previous analyses of available cellular networks, including protein-protein interaction networks, transcriptional regulatory interaction networks and metabolic interaction networks, display disassortative topologies in which the connections among hub genes are systematically suppressed, and the high connectivity for the hub genes in these networks are caused by enriched interactions between hub and non-hub genes (30, 31).

Figure 3.4. Scale-free characteristic in the genetic architecture of complex traits. Distribution of connectivity of epistatic interactions for sensitivity to heat shock (37 ^{o}C) (**A**), benomyl (**B**) and NaCl (**C**). The *p* values for fitting to power-law distribution are shown. (**D**) Cumulative distribution of all *p* values for fitting connectivity of epistatic interactions to power-law distribution in each of the 354 biological traits. In 98% of all traits, the *p* values are smaller than 0.05 (indicated by the blue line). Even with bonferroni multiple test correction, ~74% of traits are still statistically significant (indicated by the orange line).

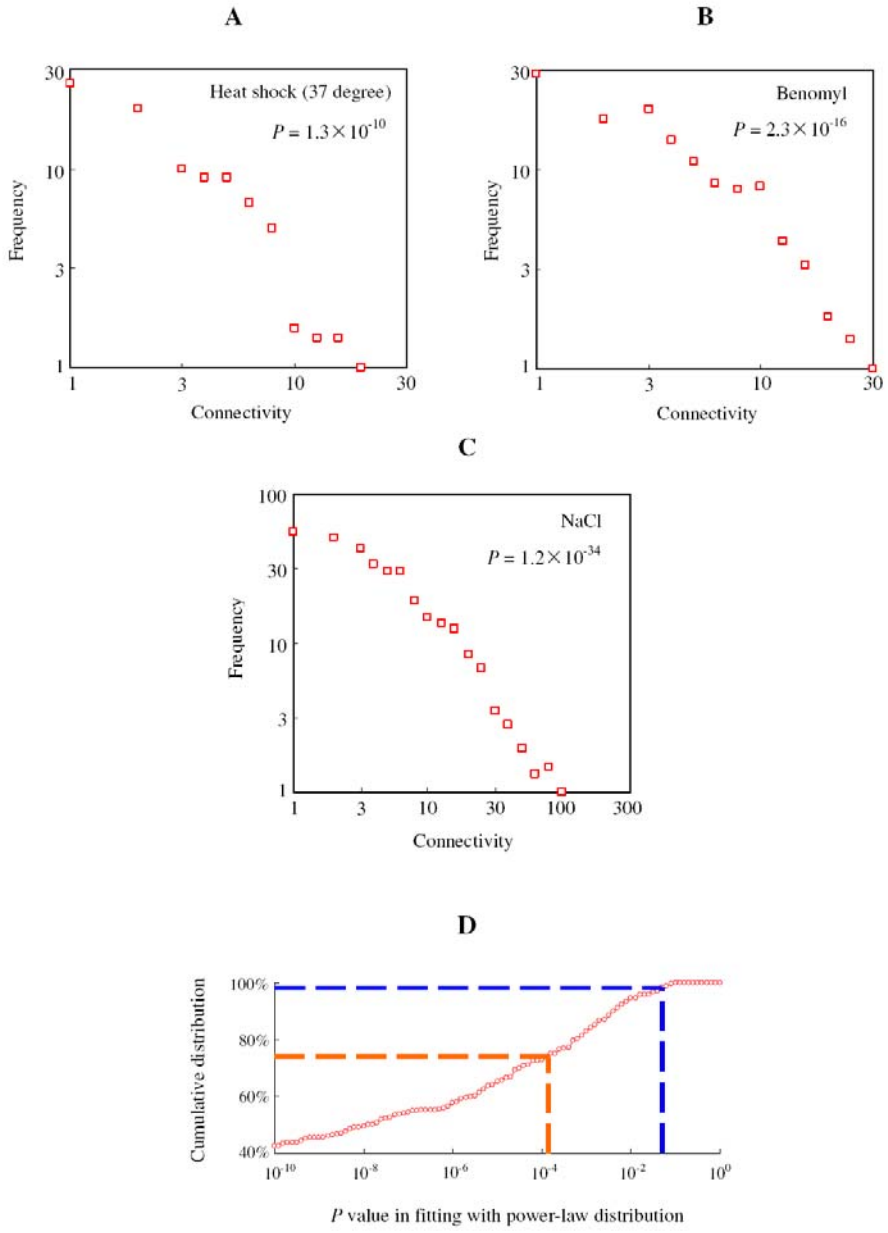
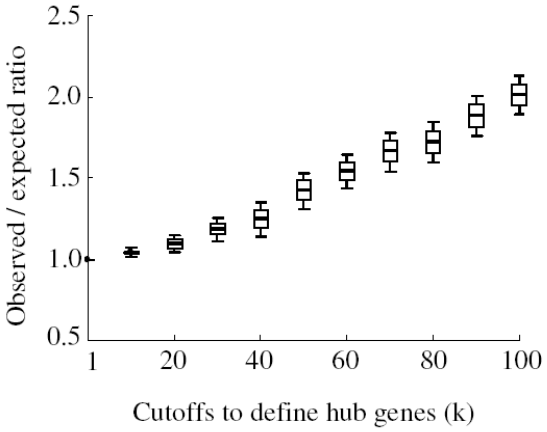
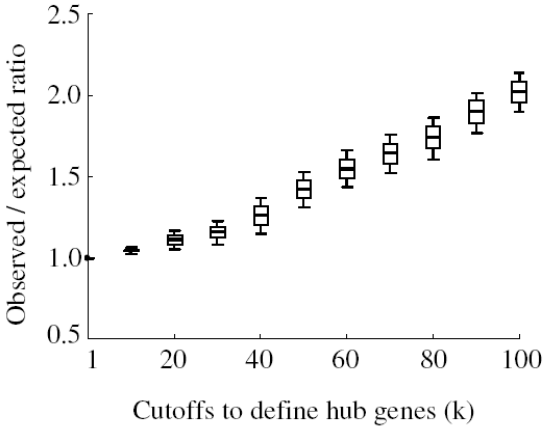

A
Heat shock (37 degree)
$P = 1.3\times10^{-10}$
Frequency
Connectivity
B
Benomyl
$P = 2.3\times10^{-16}$
Frequency
Connectivity
C
NaCl
$P = 1.2\times10^{-34}$
Frequency
Connectivity
D
Cumulative distribution
P value in fitting with power-law distribution

Figure 3.5. Assortative genetic architecture of growth traits. The average ratio of observed/expected number of epistatic interactions among 354 traits after excluding ribosomal proteins (top) and chaperones (bottom), respectively. The same simulations were conducted as in Fig. 2.3B. The bands, boxes and whiskers represent the means, ± 1 standard errors and ± 95% confidence intervals, respectively.

Removing ribosomal proteins



Removing chaperones



3. 3. 3. Implication for pleiotropy, epistasis and complex traits

Why do highly connected hub genes tend to epistatically interact with each other more frequently than expected? It was shown that the hub genes in the epistatic networks, when mutated, tend to display an impact on more phenotypes than the non-hub genes, and thus are more likely to have a higher level of pleiotropy (26). Highly pleiotropic genes would have higher chances of developing functional overlaps among themselves in the fixed functional space of a cellular system. In addition, previous studies showed that two genes with overlapping functions do tend to be linked by epistatic interactions (26, 28). As a result, highly pleiotropic hub genes would have higher chance to develop epistatic interactions among themselves. Our observation in Figure 2.3B is consistent with this scenario, indicating that pleiotropy might play an important role in shaping the genetic architecture of complex traits (32).

Although we found several novel characteristics for the genetic architecture of growth traits, several caveats need to be addressed. First, epistatic interactions used here were inferred from high-throughput experiments, which were mostly based on double gene deletion mutants. These mutations are likely to be different from most epistatically interacting mutations that underlie organism phenotypic differences in nature. Second, growth under environmental perturbations was used to represent biological traits (25), which are also different from naturally occurring phenotypic traits. Third, the epistatic interactions we used were measured only under one experimental condition (26); real epistasis underlying the growth traits might occur in different environmental conditions. Fourth, the epistatic interactions which are deduced from single and

double mutants are incomplete, because real epistasis underlying the growth traits could exist among more than two genes. Further high-throughput dissections, if possible, on the phenotypic consequences of naturally occurring genetic variations will help illustrate the genetic architecture of growth traits. For the moment, however, the approach in this study, which was used in recent studies (e.g., ref. 33), represent excellent tools to investigate this issue. With these limitations in mind, our observations identified several important features of the genetic architecture of growth traits, and indicate the importance of future efforts for addressing the architecture of epistatic interaction networks in illustrating the genetic basis of complex traits, including human diseases.

3. 4. Methods

Data resource

This study is mainly based on the integration of two high-throughput experimental datasets: a genome-wide screen for the fitness effects of gene deletion mutants under 354 conditions in *S. cerevisiae* (25) and a global survey for the epistatic interactions among more than five million gene pairs in *S. cerevisiae* (26).

In Hillenmeyer et al (2008) (25), ~6,000 heterozygous gene deletion mutants were screened in a total of 354 unique conditions (e.g. drugs approved by the U.S. Food and Drug Administration, well-characterized chemical probes, and compounds with uncertain biological activity). Genes whose heterozygous deletions significantly affect organism growth in a specific condition were defined as genes that contribute to organism growth in that condition. The authors defined significant growth defect with correction for multiple comparisons by controlling the false discovery rate (FDR) to 0.1. Growth conditions with the same chemical compound but different concentrations were regarded as the same condition, and all genes identified in different concentrations of the same compound were regarded as contributing genes under the condition. In average there are 368 genes in each sub-network, and all related data were downloaded from http://chemogenomics.stanford.edu:16080/supplements/global/download.html.

In the original synthetic genetic array (SGA) study (26), the authors screened 1,712 *S. cerevisiae* query genes, including 334 conditional or hypomorphic alleles of essential genes, against 3,885

array genes to generate a total of more than five million gene pairs spanning all biological processes. These queries were selected randomly with respect to function, while the array genes represented the whole collection of nonessential genes. In each gene pair, the epistasis value is calculated based on the equation: $\varepsilon=W_{xy}-W_xW_y$, in which W_{xy} is the fitness of an organism with mutations in both genes *X* and *Y*, whereas W_x refers to the organism with the mutation in gene *X* but not gene *Y* (and vice versa for W_y). In addition, a statistical confidence measure (p-value) was assigned to each interaction based on a combination of the observed variation of each double mutant across four experimental replicates and estimates of the background log-normal error distributions for the corresponding query and array mutants. Finally, a defined confidence threshold ($|\varepsilon| > 0.08$, $P < 0.05$) was applied to identify epistatic interactions in ref. 26. The gene pairs with epistatic interactions were downloaded from http://drygin.ccbr.utoronto.ca/~costanzo2009/.

Calculation of clustering coefficient for epistatic sub-networks

The clustering coefficient is a measure of the degree to which nodes in a network tend to be clustered together. For the node *j* with the connectivity *i* (*i*>1) in a network, its clustering coefficient C_j is defined as the following:

$$C_j = \frac{2n_j}{i(i-1)}$$

where n_j is the total number of links connecting all the neighbors of the node *j* (29). The average clustering coefficients for each of the 354 studied traits were calculated using clustering

coefficients of contributing genes in the corresponding epistasis sub-networks (34).

Statistical fitting for the scale-free distribution

Scale-free topology means that the distribution of degree in the network, *P(K)*, approximates a power law:

$$P(K)=K^{-v},$$

where K is the degree and v is the degree exponent, which is usually a constant for a specific network (29). In our analyses, the degree (K) was calculated as the number of epistatic interactions for each contributing gene in each of the 354 epistasis sub-networks. We then calculated the average frequency of each degree value among all 354 traits and plotted the frequency distribution of the network degree in Figure 2.3A. MATLAB (Mathworks) was used to fit the regression.

References

1. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet.* 5:618–625.

2. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era: concepts and misconceptions. *Nat Rev Genet.* 9:255–266.

3. Manolio TA, Collins FS, Cox NJ, et al. (27 co-authors) (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753.

4. Legare ME, Bartlett FS, Frankel WN (2000) A Major effect QTL determined by multiple genes in epileptic EL mice. *Genome Res.* 10:42–48.

5. Phillips PC (2008) Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 9:855–867.

6. Kondrashov AS (1982) Selection against harmful mutations in large sexual and asexual populations. *Genet Res.* 40:325–332.

7. Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL (2006) Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* 440:87–90.

8. Otto SP (2007) Unraveling the evolutionary advantage of sex. *Genet Res.* 89:447–449.

9. Presgraves DC (2007) Speciation genetics: epistasis, conflict and the origin of species. *Curr Biol.* 17:R125 – R127.

10. Hansen TF, Wagner GP (2001) Epistasis and the mutation load: a measurement theoretical approach. *Genetics* 158:477–485.

11. Kondrashov AS, Crow JF (1991) Haploidy or diploidy: which is better? *Nature* 351:314–315.

12. Musso G, Costanzo M, Huangfu M, et al. (11 co-authors) (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18:1092–1099.

13. Perez-Figueroa A, Caballero A, Garcia-Dorado A, Lopez-Fanjul C (2009) The action of purifying selection, mutation and drift on fitness epistatic systems. *Genetics* 1831:299 – 313.

14. Sanjuan R, Nebot MR (2008) A network model for the correlation between epistasis and

genomic complexity. *PLoS One* 37:e2663.

15. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I (2009) Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet.* 5:e1000578.

16. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends in Genet.* 8:312–316.

17. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. *Science* 291:1001–1004.

18. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol.* 23:561–566.

19. Ma X, Tarone AM, Li W (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS One* 3:e1922.

20. Brady A, Maxwell K, Daniels N, Cowen LJ (2009) Fault tolerance in protein interaction networks: stable bipartite subgraphs and redundant pathways. *PLoS One* 4:e5364.

21. Remold SK, Lenski RE (2004) Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. *Nat Genet.* 36:423–426.

22. Carlborg O, Jacobsson L, Ahgren P, Siegel P, Andersson L (2006) Epistasis and the release of genetic variation during long-term selection. *Nat Genet.* 38:418–420.

23. Ehrenreich IM, Stafford PA, Purugganan MD (2007) The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. *Genetics* 176:1223–1236.

24. Shao H, Burrage LC, Sinasac DS, et al. (14 co-authors) (2008) Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proc Natl Acad Sci* U S A. 105:19910–19914.

25. Hillenmeyer ME, Fung E, Wildenhain J, et al. (14 co-authors) (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320:362–365.

26. Costanzo M, Baryshnikova A, Bellay J, et al. (53 co-authors) (2010) The genetic landscape of a cell. *Science* 327:425–431.

27. Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common

human diseases. *Human Heredity* 56:73–82.

28. Tong AH, Lesage G, Bader GD, et al. (50 co-authors) (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808–813.

29. Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 5:101–113.

30. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett.* 89:208701.

31. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910–913.

32. Wagner GP, Zhang J (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex adaptations. *Nat Rev Genet.* 12:204–213.

33. Dowell RD, Ryan O, Jansen A, et al. (16 co-authors) (2010) Genotype to phenotype: a complex problem. *Science* 328:469.

34. Li J, Min R, Vizeacoumar FJ, Jin K, Xin X, Zhang Z (2010) Exploiting the determinants of stochastic gene expression in *Saccharomyces cerevisiae* for genome-wide prediction of expression noise. *Proc Natl Acad Sci* U S A. 107:10472–10477.

CHAPTER FOUR

Dynamic Epistasis for Different Alleles of the Same Gene

4. 1. Summary

Epistasis refers to the phenomenon that phenotypic consequences caused by mutation of one gene depend on one or more mutations at another gene. Epistasis is critical for understanding many genetic and evolutionary processes, including pathway organization, evolution of sexual reproduction, mutational load, ploidy, genomic complexity, speciation and the origin of life. Nevertheless, current understandings for the genome-wide distribution of epistasis are mostly inferred from interactions among one mutant type per gene, whereas how epistatic interaction partners change dynamically for different mutant alleles of the same gene is largely unknown. Here we address this issue by combining predictions from Flux Balance Analysis (FBA) and data from a recently published high-throughput experiment. Our results showed that different alleles can epistatically interact with very different gene sets. Furthermore, between two random mutant alleles of the same gene, the chance for the allele with more severe mutational consequence to develop a higher percentage of negative epistasis than the other allele is 50%~70% in eukaryotic organisms, but only 20~30% in prokaryotes and archaea. We developed a population genetics model which predicts that the observed distribution for the sign of epistasis can speed up the process of purging deleterious mutations in eukaryotic organisms. Our simulation results indicate that the epistasis among genes can be dynamically rewired at the genome level, and call on future high-throughput experimental effort to illustrate the distribution of epistasis for various alleles of the same gene.

(**Lin Xu and Brandon Barker have contributed equally to the work**)

4. 2. Introduction

Epistasis between two deleterious mutations is positive when a double mutant causes a weaker mutational defect than predicted from individual deleterious mutations, and is negative when the double mutant causes a stronger defect (1, 2). In a population with sexual reproduction, positive epistasis alleviates the total harm when multiple deleterious mutations combine together and thus reduces the effectiveness of natural selection in removing these deleterious mutations, whereas negative epistasis can lower average mutational load by efficiently purging deleterious mutants (3). As a consequence, selective elimination of deleterious mutations would be especially effective if negative epistasis is prevalent. It is important to understand the distribution of epistasis among mutations, which plays a central role in genetics and theoretical descriptions for many evolutionary processes (reviewed in refs. 1 and 2).

Tremendous efforts have been put into genome-wide measurements for the sign and magnitude of epistasis among different genes in various species (4-15). A series of high-throughput experimental platforms have been developed, such as synthetic genetic array (SGA) (4, 5), diploid-based synthetic lethality analysis with microarrays (dSLAM) (6, 7), synthetic dosage-suppression and lethality screen (8-10) and epistatic miniarray profiles (EMAP) (11-13). The epistatic relations in these experiments were mostly measured based on one mutant type (deletion mutant) per gene. Few studies constructed multiple mutant alleles for single genes to examine the dynamics of epistatic relations among genes under different genetic perturbations. As a consequence, the global landscape of epistasis for different alleles of the same gene remains

largely uninvestigated.

Here we address this issue by combining experimental data with mathematical modeling using Flux Balance Analysis (FBA). FBA involves the optimization of cellular objective functions and allows prediction of *in silico* flux values and/or growth (16-18). FBA has been used to investigate the fitness consequence of single deletion mutants (19, 20) and epistatic relations between metabolic reactions, genes and functional modules (21-24). The FBA predictions show good agreement with genome-wide experimental studies (25-32). One essential advantage of FBA modeling is that it can simulate epistasis between genes based on different genetic mutants. Using this platform, together with data from a recently published experiment, we were able to show that epistasis can be rewired among genes and that the sign of epistasis can change dramatically at the global scale, depending on mutant alleles involved in the processes. Our study provides the first genome-wide picture on the dynamic epistatic landscape of various mutant alleles for the same gene.

4. 3. Results

4. 3. 1. Epistatic relations between genes are largely allele-specific

We first used the yeast *S. cerevisiae* metabolic reconstruction iMM904 (16) to examine the distribution of epistasis under various genetic mutant alleles. The reconstruction is a genome-scale metabolic model, having 904 metabolic genes associated with 1,412 reactions. For each gene, we simulated genetic perturbations that retain the corresponding flux from 90% (slightly deleterious mutation) to 0% (severely deleterious mutations) in decrement of 10% of its wild-type (optimal) flux. As a result, ten different single mutants per non-essential gene and nine different single mutants per essential gene (the 0% flux mutants in these genes represent lethal deletion for which epistasis cannot be calculated) were simulated. We computed the fitness of the single mutants and double mutants with any possible pairwise allele combination of different genes. These data were used to infer the epistatic relationships among genes in the genome. In total, over 40 million simulations were conducted.

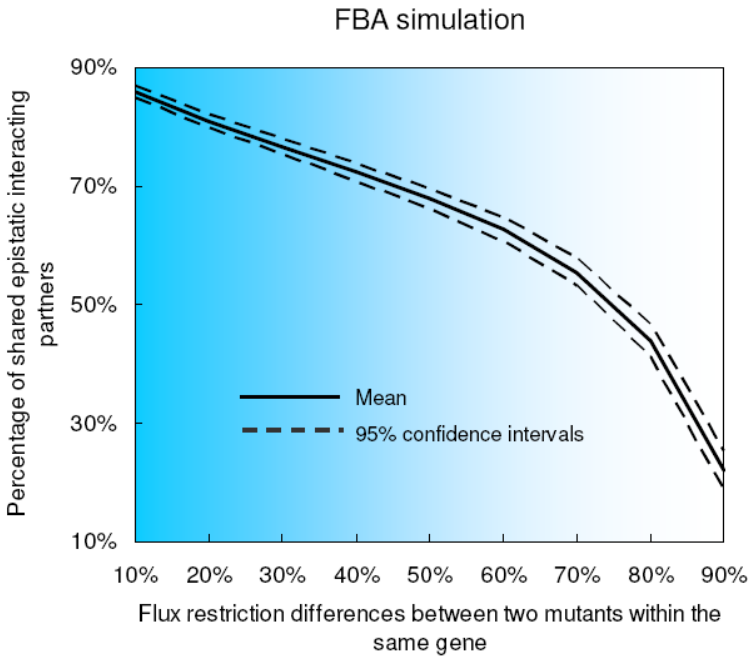
To investigate the dynamics of epistasis among genes, we calculated the percentage of shared epistatic interaction partners between any two mutants within the same gene. Two mutants are defined to share an epistatic interaction partner (a mutant from another gene) if they both epistatically interact with this mutant and the sign of epistasis are the same. The percentage of shared epistatic interaction partners between two mutants is calculated as the number of their shared epistatic interaction partners divided by the sum of their total epistatic interaction partners. As shown in Fig. 4.1, our results indicate that the percentage of shared epistatic interaction

partners between two mutants of the same gene decreases as the flux difference between them increases. Two mutants of the same genes could have as low as only ~20% overlap between their epistatic interaction partners, indicating that epistatic profile of a gene are largely dependent on the mutant types used. We also repeated the above FBA analysis for another species, *Escherichia coli*, and our results confirmed the above trend (Figure 4.2).

In a recently released high-throughput experiment that measured genome-wide epistatic relations among genes in *S. cerevisiae* (4), there were 43 genes having two different mutant alleles, each of which were experimentally crossed with 3,885 array gene deletion mutants to explore their epistatic relations in the genome. This dataset provides the most comprehensive experimental source for investigating the epistatic landscape of different mutant alleles in the same gene. Fig. 4.1B shows the empirical cumulative distribution for the percentage of shared interaction partners between mutant pairs within the same gene. Our results indicate that more than 50% of mutant pairs within the same gene have less than 10% overlap of their epistatic interaction partners, and ~90% mutant pairs have less than 20% overlap (Fig. 4.1B). Although the genes used in the experiments are totally different from the ones in the FBA model, the result confirms FBA modeling prediction that different mutant alleles of the same gene can have very distinct epistatic interaction partners in the genome.

Figure 4.1. The profiling of epistatic interaction partners for each gene is largely dependent on mutant types involved. **(A)** FBA simulation results for the distribution of the percentage of shared epistatic interaction partners between two mutant alleles within the same gene. Solid and broken lines represent mean and 95% confidence intervals, respectively. **(B)** The cumulative distribution for the percentage of shared epistatic interaction partners between two mutant alleles within the same gene based on real experimental data. Two broken lines represent 10% and 20% of shared epistatic profiling, respectively.

A


FBA simulation
Percentage of shared epistatic interacting partners
90%
70%
50%
30%
10%
Mean
95% confidence intervals
10% 20% 30% 40% 50% 60% 70% 80% 90%
Flux restriction differences between two mutants within the same gene

B

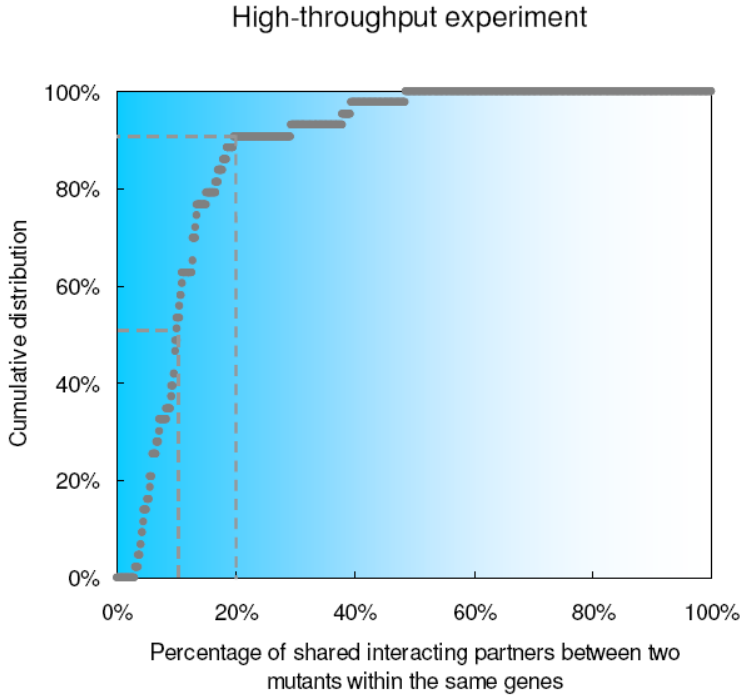
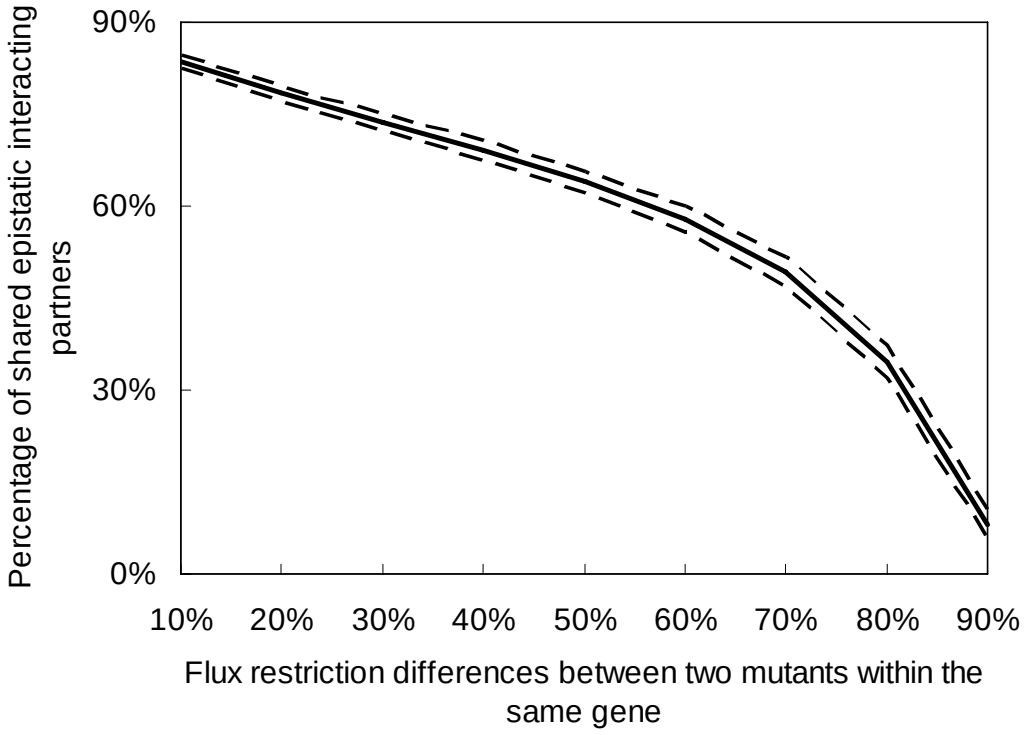

High-throughput experiment
Cumulative distribution
100%
80%
60%
40%
20%
0%
0% 20% 40% 60% 80% 100%
Percentage of shared interacting partners between two mutants within the same genes

Figure 4.2. Percentage of shared epistatic interacting partners based on flux differences between two mutant alleles of the same gene. The analysis procedure is the same as Fig. 1A, but instead of using the *S. cerevisiae* model, here we repeated the analysis using the *E. coli* model (54).

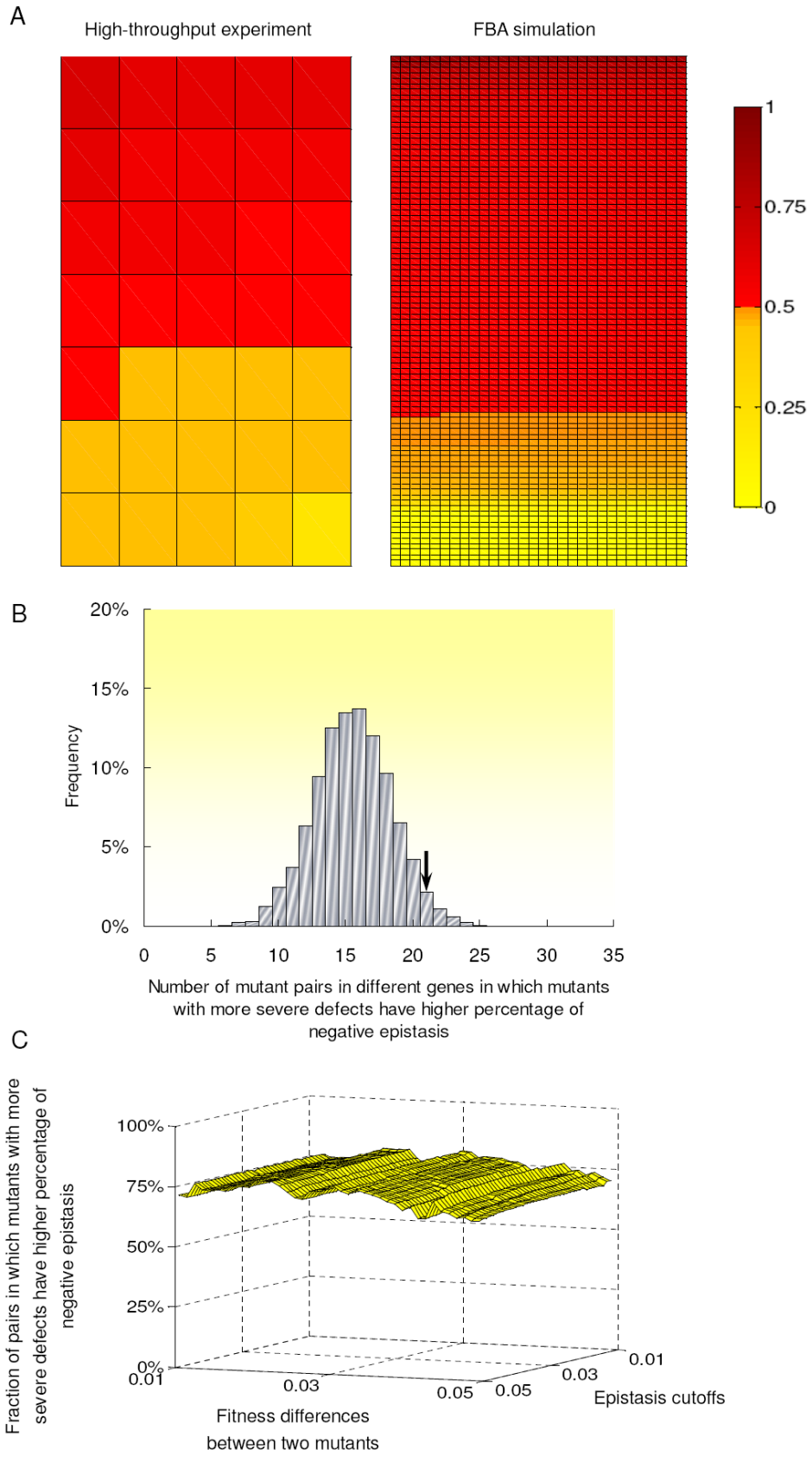

Percentage of shared epistatic interacting partners
90%
60%
30%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Flux restriction differences between two mutants within the same gene

4. 3. 2. The sign of epistasis for individual gene depends on mutation severity

The relative prevalence of positive versus negative epistasis is of tremendous importance for understanding many evolutionary processes (1-3). In the following we addressed this issue for different alleles of the same gene. Based on the above high-throughput experimental dataset, we calculated the percentage of negative epistasis for each mutant, defined as the number of the mutant's negative epistatic partners divided by the overall number of its epistatic partners. We then compared the percentage of negative epistasis between different mutant alleles of the same gene in the experiment. Among 43 mutant pairs in the study, 35 mutant pairs have significantly different growth (fitness) between two mutants of the same gene. As shown in Fig. 4.3A (left panel), 21 mutant pairs (60%) show that alleles with more severe defects in the same gene have a higher chance than the other allele to develop negative epistasis in the genome.

To see if this result could be caused by possible systematic trend in the high-throughput experiments, we randomly selected 35 pairs of gene deletion mutants that have the same growth-rate difference and compared their relative prevalence of negative epistasis. The permutation was repeated 100,000 times and the result is depicted in Fig. 4.3B. Among each repeat of randomly selected 35 mutant pairs, only a small percentage (4.2%) have 21 or more mutant pairs where the mutant with more severe defects has a higher chance than the other mutant to develop negative epistasis in the genome, indicating that our observation for different mutant alleles of the same gene is not likely caused by the overall pattern in the high-throughput experiments.

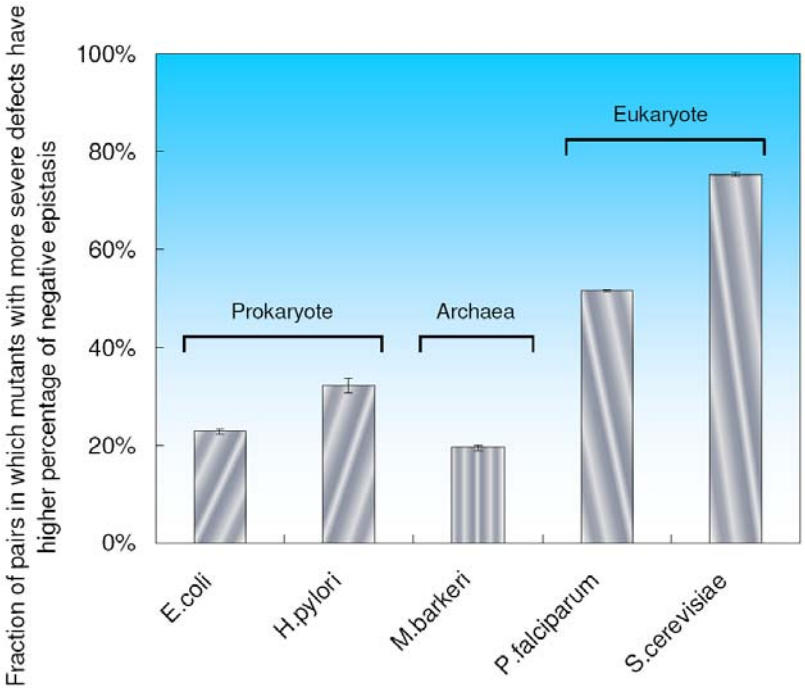
Figure 4.3. Mutant alleles in the same gene with more severe defects tend to have a higher percentage of negative epistasis in yeast. **(A)** The two matrices represent all mutant pairs identified in real experimental data (left panel) and FBA simulation (right panel) (fitness difference $|f| \geq 0.05$, epistasis cutoff $|\varepsilon| \geq 0.01$). Each cell represents one mutant pair within the same gene. The colorbar to the right represents the normalized percentage of negative epistasis for the mutant allele with more severe defects (percentage of negative epistasis for the mutant allele with more severe defects / the sum of percentage of negative epistasis for two mutant alleles). Red and yellow colors represent that mutant allele with more severe defects in the same gene has higher and lower percentage of negative epistasis than the other allele, respectively. (B) Distribution for the number of mutant pairs among randomly selected 35 pairs where mutants with more severe defects have higher percentage of negative epistasis. The arrow represents the observed number for the mutant allele pairs within the same genes. **(C)** The percentage of mutant pairs in which mutant allele with more severe defects in the same gene has higher percentage of negative epistasis under various fitness difference and epistasis cutoffs during FBA simulations.


A
High-throughput experiment
FBA simulation
1
0.75
0.5
0.25
0
B
20%
15%
10%
5%
0%
Frequency
0
5
10
15
20
25
30
35
Number of mutant pairs in different genes in which mutants with more severe defects have higher percentage of negative epistasis
C
Fraction of pairs in which mutants with more severe defects have higher percentage of negative epistasis
100%
75%
50%
25%
0%
0.01
0.03
0.05
Fitness differences between two mutants
0.01
0.03
0.05
Epistasis cutoffs

Using results from the above FBA simulation, we also confirmed the same pattern that mutant alleles with more severe defects in the same genes have a higher chance to develop negative epistasis with other mutants in the yeast genome (Fig. 4.3A, right panel). Indeed, an even higher percentage of mutant allele pairs in the FBA simulation (~70%) than in real experiments (60%) support this conclusion. To avoid possible bias from the definition of epistasis and fitness differences between mutant alleles in FBA simulation, we repeated the calculations based on multiple criteria and our conclusion remains the same (Fig. 4.3C).

Our observation that mutant alleles with more severe defects in the same gene have a higher chance to develop negative epistasis is surprising, given previous results based on virus models or gene network simulations (33-37) which proposed a totally opposite pattern at the genome level, *i.e.,* mutations with larger mutational defects are more likely to develop positive epistasis. We further used the FBA simulations to explore the dynamics of epistasis under various mutant alleles in different species. High quality genome-wide metabolic networks in two prokaryotes (*E. coli* (38), *H. pylori* (39)), one archaea (*M. barkeri* (40)) and another single cell eukaryote (*P. falciparum* (41)) were used in our simulation. As shown in Fig. 4.4, when two mutant alleles of the same gene are compared, in 22%, 32%, and 19% cases for *E. coli*, *H. pylori* and *M. barkeri*, respectively, mutant alleles with more severe defects display higher percentages of negative epistasis than the other allele, indicating that more deleterious mutant alleles in the same gene indeed tend to develop positive epistasis in these species. However, these numbers are significantly smaller than that of another eukaryotic organism, *P. falciparum* (52%).

Figure 4.4. Mutant alleles with more severe defects tend to have a higher percentage of negative epistasis in eukaryotes than prokaryotes and archaea. The Y axis shows the percentage of mutant pairs in which mutant alleles with more severe defects in the same gene has higher percentage of negative epistasis than the other allele. FBA simulations were conducted for two prokaryote species (*E. coli* and *H. pylori*), one archaea species (*M. barkeri*) and two single-cell eukaryote species (*P. falciparum* and *S. cerevisiae*). The mean and standard errors were based on results from 40 epistasis cutoff values ranging from 0.01 to 0.05.


Fraction of pairs in which mutants with more severe defects have higher percentage of negative epistasis
100%
80%
60%
40%
20%
0%
Prokaryote
Archaea
Eukaryote
E.coli
H.pylori
M.barkeri
P.falciparum
S.cerevisiae

4. 3. 3. A self-purging mechanism for deleterious mutations

Our above results indicate that between two random mutant alleles of the same gene, the chance for the allele with more severe mutational consequence to develop a higher percentage of negative epistasis than the other allele is 50%~70% in eukaryotic organisms, but only 20~30% in prokaryotes and archaea. In other words, mutant alleles with more severe defects in the same gene might have a higher chance to develop negative epistasis in eukaryotic organisms than in prokaryotes and archaea. We constructed a simple population genetic model as in Fig. 4.5A to address the evolutionary significance of this observation. The genetic system has two genes: a query gene A which contains three different alleles (A^S: mutants with severe defects; A^D: mutants with weak defects; A^{WT}: wild type), and a gene X which has two different alleles (mutant, X^M and wild type, X^{WT}). We simulated the dynamics of allele frequency between the severe and weak mutant alleles in the gene A under different probabilities of having negative epistasis between these two alleles and the mutant allele in the gene X. Our results in Fig. 4.5B depict the simulation results. The six panels in the figure represent the ratio of allele frequency of the severe to the weak mutant alleles for the gene A in the 50th, 100th, 150th, 200th, 250th and 300th generations, respectively. Our simulations indicate that if the percentage of negative epistasis for the severe mutant is kept as a constant, as the percentage of negative epistasis for the weak mutation increases (as shown by the arrow A), the ratio of the severe to the weak allele frequency would increase. However, this ratio would decrease, indicating a faster removal of the severe mutants from the population, in another direction (as shown by the arrow B), *i.e*., the percentage of negative epistasis for the weak mutant is kept as a constant, but the percentage of negative

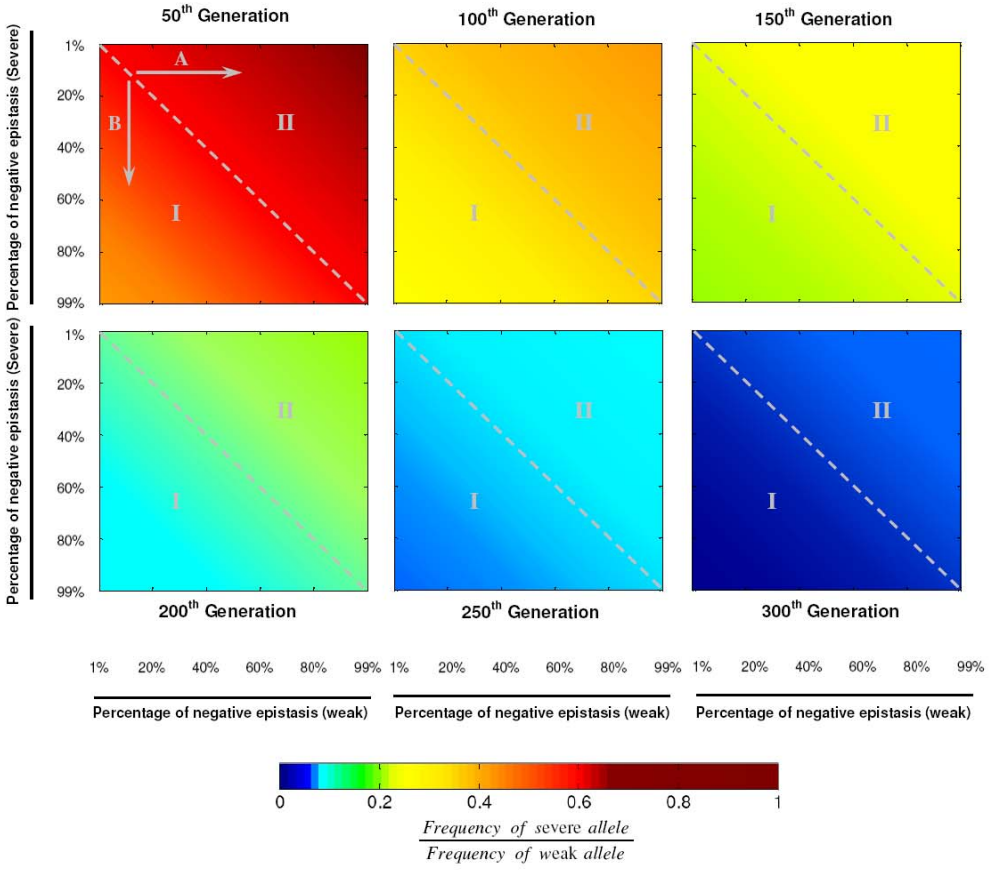
Figure 4.5. Increased efficiency of purging deleterious mutations in eukaryotic organisms. **(A)** The population genetics model for allele frequency changes from generation to generation. In the figure, p and ω represent allele frequency and fitness, respectively. A and X are genes with different alleles, and ε is the epistasis term between mutant types of different genes. **(B)** The ratio of the severe to the weak alleles in the 50th, 100th, 150th, 200th, 250th and 300th generations. Colors represent the ratio as indicated at the bottom. The diagonal line in each panel represents the situation where the severe and the weak mutant alleles have the same probability of having negative epistasis in the genome. It is noteworthy to point out that in each panel the ratio of the severe to the weak alleles increases from the bottom-left (Region I, the severe mutant has more negative epistasis) to the upper-right (Region II, the weak mutant has more negative epistasis) part of the panel. The arrows A and B are discussed in the text.

A

Genotype	Before selection (Generation T) Frequency	Relative fitness	After selection (Generation T+1) Frequency
$\begin{bmatrix} A^S X^M & A^S X^{WT} \\ A^D X^M & A^D X^{WT} \\ A^{WT} X^M & A^{WT} X^{WT} \end{bmatrix}$	$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \end{bmatrix}$	$\begin{bmatrix} \omega_{11} = \omega_A^S \times \omega_X^M + \varepsilon & \omega_{12} = \omega_A^S \\ \omega_{21} = \omega_A^D \times \omega_X^M + \varepsilon & \omega_{22} = \omega_A^D \\ \omega_{31} = \omega_X^M & \omega_{32} = 1 \end{bmatrix}$	$\begin{bmatrix} \frac{p_{11}\omega_{11}}{\bar{\omega}} & \frac{p_{12}\omega_{12}}{\bar{\omega}} \\ \frac{p_{21}\omega_{21}}{\bar{\omega}} & \frac{p_{22}\omega_{22}}{\bar{\omega}} \\ \frac{p_{31}\omega_{31}}{\bar{\omega}} & \frac{p_{32}\omega_{32}}{\bar{\omega}} \end{bmatrix}$

where $\bar{\omega} = p_{11}\omega_{11} + p_{12}\omega_{12} + p_{21}\omega_{21} + p_{22}\omega_{22} + p_{31}\omega_{31} + p_{32}\omega_{32}$

B



epistasis for the severe mutant increases. Therefore, the distribution for the sign of epistasis among different alleles of the same gene observed in this study might represent an efficient way for eukaryotic organisms to purge out deleterious mutations from populations.

4. 4. Discussion

Our study represents the first theoretical survey for the dynamics of global epistatic effects under various mutant alleles of the same gene. We show that epistatic profiling of a gene at the genome level is largely dependent on mutant types involved. Our results indicate that previous conclusions inferring epistatic relations among genes based on only one mutant type per gene can be greatly improved by using multiple mutant alleles. More importantly, our study for the first time show that mutant alleles with severe defects have a higher chance to develop negative epistasis in eukaryotic organisms than in prokaryotes and archaea. It has been speculated that eukaryotic organisms might have more negative epistasis due to their increased complexity over prokaryotic organisms (42, 43). Even if this hypothesis is true, however, our results for different mutant alleles of the same gene cannot be directly inferred from this complexity argument. The real mechanism underlying our observation remains to be determined.

The origin and maintenance of sexual reproduction remains one of the central issues in evolutionary biology. The mutational deterministic hypothesis for the maintenance of sexual reproduction posits that sex enhances the ability of natural selection to purge deleterious mutations after recombination brings them together into single genome (44). This explanation requires the prevalence of negative epistasis at the genome level. Here we found that the mutations with larger deleterious defects within the same gene have a higher chance to develop negative epistasis in eukaryotic organisms than prokaryotes and archaea. Our simple population genetics model indicates that this novel distribution of negative epistasis among different alleles

of the same gene in eukaryotic organisms might be able to efficiently purge deleterious mutations from the populations, thus providing a previously unappreciated evolutionary advantage for sexual reproduction. We have to emphasize that these findings do not necessarily provide sufficient evidence to explain the cause for the emergence of sexual reproduction during evolution.

Although we found several novel characteristics regarding the global epistatic landscape of different mutant alleles in the same gene, two caveats need to be addressed. First, the FBA modeling used in this study, which has been successfully applied to various research problems (19-24), only includes metabolic genes in the simulation. However, results from our analysis on the epistatic relations among ~0.2 million double mutants comprising of ~4,000 *S. cerevisiae* genes, which nearly represent all functional categories in the budding yeast, confirmed the major FBA modeling predictions and therefore indicate that the model could be used to examine the global trend of epistatic effects. Second, even though FBA is one of the most comprehensive computational tools for simulating epistatic interactions among genes, there are still many aspects that can be improved to aid in capturing the full set of empirical genetic interactions (45). The quality and completeness of the metabolic reconstructions would be one of the main reasons for the inconsistency between gene-specific model predictions and the empirical data. In addition, lack of information from transcriptome and interactome would also be responsible for the limitations of predictive power in FBA. In the future, we shall integrate rules for transcriptional regulation and physical interactions into this framework to improve over the current FBA

methods in predicting epistasis (46). With these limitations in mind, our observations identified several important features for the dynamics of epistasis among genes, and call on future experimental effort to examine the distribution of epistasis with a wide array of mutants representing various allele types of the same gene.

4. 5. Methods

Experimental dataset

The experimental data was extracted from a global survey for the epistatic interactions among more than five million gene pairs in *S. cerevisiae* (4). In the original synthetic genetic array (SGA) study (4), the authors screened 1,712 *S. cerevisiae* query gene mutants against 3,885 array gene mutants to generate a total of more than five million gene pairs spanning all biological processes. In each gene pair, the epistasis value is calculated based on the equation: $\varepsilon=W_{xy}-W_xW_y$, in which W_{xy} is the fitness of an organism with two mutations in genes *X* and *Y*, whereas W_x or W_y refers to the fitness of organism with mutation only at gene *X* or *Y,* respectively. In addition, a statistical confidence measure (*P-value*) was assigned to each interaction based on a combination of the observed variation of each double mutant across four experimental replicates and estimates of the background log-normal error distributions for the corresponding query and array mutants. Finally, a defined confidence threshold ($|\varepsilon| \geq 0.01$, $P < 0.05$) is applied to generate epistatic interactions.

Population genetics model

We simulated a deterministic genetic system with a query gene A, which contains three different alleles (A^S, severe mutant, A^D, weakly deleterious mutants, and A^{WT}, wild type), and a gene X that has two different alleles (X^M, mutant and X^{WT}, wild type). Different alleles of gene A could meet with alleles in gene X from generation to generation. The mutant alleles of genes A and X could have three distinct epistatic relations: positive epistasis, negative epistasis and no epistasis.

We simulated the ratio of allele frequency between the severe and the weak mutant alleles under all possible percentages of negative epistasis of A^S and A^D alleles, by randomly choosing above three epistatic relations for A^S - X^M and A^D - X^M pairs separately in each of 1,000 simulations.

The table in Fig. 4.5A explains how allele frequency of various mutations could be calculated from generation T to generation T+1 under natural selection. In the figure, p and ω represent allele frequency and fitness, respectively. The average fitness in generation T could be calculated based on the following equation (47):

$$\overline{\omega} = p_{11}\omega_{11} + p_{12}\omega_{12} + p_{21}\omega_{21} + p_{22}\omega_{22} + p_{31}\omega_{31} + p_{32}\omega_{32}$$

To make the simulation simple, the initial allele frequencies for the severe, weak and WT alleles for the A gene are assumed to be equal (1/3). In addition, the initial allele frequencies for the mutant and WT for X gene are also assumed to be equal (1/2). The fitness difference between severe and weak alleles of the gene A is 1% in the simulation. The epistasis values (ε) between the A and X gene mutants are constant in the simulation (0.01 and -0.01 for positive and negative epistasis, respectively). We assumed that 10% of double mutants display epistatic relations in the simulation, which is consistent with the known experimental evidences (4). The average ratio of allele frequency of the severe to the weak alleles is calculated for different levels of negative epistasis occurrence (Fig. 4.5B). A variety of fitness differences between the severe and weak alleles and epistasis values have been tested and the trend remains the same (data not shown).

References

1. Phillips PC (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9:855-867.

2. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet* 8:437-449.

3. Kimura M, Maruyama T (1966) The mutational load with epistatic gene interactions in fitness. *Genetics* 54:1337–1351.

4. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327:425-431.

5. Tong AH, et al. (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808-813.

6. Pan X, et al. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol Cell* 16:487-496.

7. Pan X, et al. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 124:1069-1081.

8. Measday V, Hieter P (2002) Synthetic dosage lethality. *Methods Enzymol* 350:316-326.

9. Measday V, et al. (2005) Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci USA* 102:13956-13961.

10. Sopko R, et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21:319-330.

11. Collins SR, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446:806-810.

12. Kornmann B, et al. (2009) An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science* 325:477-481.

13. Fiedler D, et al. (2009) Functional Organization of the *S. cerevisiae* Phosphorylation Network. *Cell* 136:952-963.

14. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for

positive epistasis in HIV-1. *Science* 306:1547-1550.

15. Roguev A, et al. (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 322:405-410.

16. Mo ML, Palsson BØ, Herrgard MJ (2009) Connecting extracellular metabolomic profiles to intracellular metabolic states in yeast. *BMC Syst Biol* 3:37.

17. Becker SA, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2:727-738.

18. Smallbone K, Simeonidis E (2009) Flux balance analysis: a geometric perspective. *J Theor Biol* 258:311-315.

19. Papp B, Pal C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429:661-664.

20. Ibarra RU, Edwards JS, Palsson BØ (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420:186-189.

21. Harrison R, Papp B, Pal C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci USA* 104:2307-2312.

22. Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* 38:993-998.

23. Segrè D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* 37:77-83.

24. He X, Qian W, Wang Z, Li Y, Zhang J (2010) Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat Genet* 42:272-276.

25. Edwards JS, Ibarra RU, Palsson BØ (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19:125-130.

26. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 99:15112-15117.

27. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 102:7695-7700.

28. AbuOun M, et al. (2009) Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J Biol Chem* 284:29480-29488.

29. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* 33:164-190.

30. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli. Nat Biotechnol* 26:659-667.

31. Fong SS, et al. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 91:643-648.

32. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res* 15:1365-1372.

33. Burch CL, Chao L (2004) Epistasis and its relationship to canalization in the RNA virus φ6. *Genetics* 167:559-567.

34. You L, Yin J (2002) Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage T7. *Genetics* 160:1273-1281.

35. Sanjuán R (2006) Quantifying antagonistic epistasis in a multifunctional RNA secondary structure of the Rous sarcoma virus. *J Gen Virol* 87:1595-1602.

36. Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL (2006) Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* 440:87-90.

37. Lohaus R, Burch CL, Azevedo RB (2010) Genetic architecture and the evolution of sex. *J Hered* 101:S142-S157.

38. Feist AM, et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.

39. Thiele I, Vo TD, Price ND, Palsson BØ (2005) Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol* 187:5818-5830.

40. Feist AM, Scholten JC, Palsson BØ, Brockman FJ, Ideker T (2006) Modeling

methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2:2006.0004.

41. Plata G, Hsiao TL, Olszewski KL, Llinás M, Vitkup D (2010) Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network. *Mol Syst Biol* 6:408.

42. Sanjuán R, Elena SF (2006) Epistasis correlates to genomic complexity. *Proc Natl Acad Sci USA* 103:14402-14405.

43. Sanjuán R, Nebot MR (2008) A network model for the correlation between epistasis and genomic complexity. *PLoS One* 37:e2663.

44. Otto SP (2007) Unraveling the evolutionary advantage of sex. *Genet Res* 89:447-449.

45. Covert MW, Schilling CH, Palsson BØ (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* 213:73-88.

46. Szappanos B, et al. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* 43:656-662.

47. Hartl DL, Clark AG (2007) *in Principles of Population Genetics*. 4 ed. Sinauer Associates, Sunderland, Massachusetts.

CHAPTER FIVE

Epistatic Landscapes under Varying Environmental Perturbations

5. 1. Summary

Epistasis refers to the phenomenon that phenotypic consequences caused by mutation of one gene depend on mutations at another gene. It is well known that the global epistatic landscape under varying environments is critical for understanding many genetic and evolutionary processes, including evolution of sexual reproduction, mutational load, ploidy, genomic complexity, speciation and the origin of life. Nevertheless, current understanding for the epistatic landscape is severely limited by the incapability to screen epistatic relations among genes under various environmental perturbations. Here we address this issue by applying Flux Balance Analysis (FBA) to simulate epistatic landscapes under various environmental perturbations. Our results showed that 28% and 24% of epistatic relations are extremely conserved (exist in all conditions) or extremely dynamic (exist in only one condition) respectively, across the 16 conditions we simulated. We also discovered that gene pairs with FBA-predicted epistatic interactions are more likely to experience co-evolution than random expectation. In addition, our results indicate that genes with extremely conserved epistasis have more similar evolutionary rates than that of extremely dynamic epistasis and that these two types of epistatic interaction networks have distinct topologies. Our findings provide a new genome-wide perspective to understand epistatic dynamics under varying environmental perturbations.

(**Lin Xu and Brandon Barker have contributed equally to the work**)

5. 2. Introduction

Epistasis refers to the phenomenon wherein mutations of two genes can modify each other's phenotypic outcomes. It is also well established that epistasis is important for the evolution of sex (1-3), speciation (4), mutational load (5), ploidy (6, 7), genetic drift (8), genomic complexity (9), drug resistance (10), and human disease (11). Here we focus on studying how environmental perturbations could influence the dynamics of epistatic networks in the biological systems.

One of the main obstacles to exploring epistatic dynamics under a variety of environments is the lack of high-throughput experimental platforms. To explore epistasis on a genomic scale, a number of technologies have been developed to systematically map genetic interaction networks, such as synthetic genetic array (SGA) (12, 13), diploid-based synthetic lethality analysis with microarrays (dSLAM) (14, 15), synthetic dosage-suppression and lethality screen (16-18) and epistatic miniarray profiles (EMAP) (19-21). A key issue for all these experiments is that these epistatic networks have been constructed only under normal laboratory conditions. However, cells are constantly bombarded by various external environmental stresses. Epistatic dynamics under these perturbations can not be predicted based on normal laboratory conditions. Few studies have constructed epistatic networks under multiple environmental perturbations, because the effort required to accomplish this aim is far beyond the capability of current experimental platforms. For example, a recent genome-wide study that has only constructed epistatic networks under one normal and one stressful conditions already requires a large amount of experimental resources (22). As a consequence, the global landscape of epistasis under a variety of

environmental perturbations remains largely uninvestigated.

Here we address this issue by using Flux Balance Analysis (FBA) to simulate epistatic dynamics under multiple environmental perturbations. FBA involves the optimization of a cellular objective function subject to the reactions and constraints of a metabolic network, which can provide reliable predictions (23-32). Using this platform, we are able to show that, while a large proportion of epistasis can be rewired dramatically under varying environments, there are still many epistatic interactions that are conserved under varying environmental perturbations. More interestingly, we find gene pairs with FBA-predicted epistasis are more likely to co-evolve than random simulations, which indicates that FBA simulations could correctly predict biological characteristics of gene pairs with epistasis. Furthermore, we find some unique network properties among extremely conserved and extremely dynamic epistatic interactions. Our study thus provides the first genome-wide picture on the dynamic epistatic landscape under multiple environmental perturbations.

5. 3. Results

5. 3. 1. Global epistatic landscape under multiple environmental perturbations

We applied the yeast *S. cerevisiae* metabolic reconstruction iMM904 (33) to examine the dynamics of epistasis under various environmental perturbations. The reconstruction is a genome-scale metabolic model, having 905 metabolic genes associated with 1,412 reactions. We conducted FBA simulations under the following 16 environmental perturbations. In 15 of these perturbations, the carbon source (abundant glucose) was replaced with one of: acetaldehyde, acetate, adenosine 3',5'-bisphosphate, adenosyl methionine, adenosine, allantoin, alanine, arginine, ethanol, glycerol, glutamine, glutamate, low glucose, trehalose, and xanthosine, respectively. Additionally, we looked at abundant glucose under limited phosphorus availability. To assure all these environmental conditions having the same growth rates in the following analyses, we restricted the carbon source or phosphorous uptake levels for each of the 16 environmental perturbations, such that only 20% of the high-glucose growth rate was attained. Epistatic relations between any two genes were calculated under each of the 16 harsh environments (see Methods).

In order to compare epistatic dynamics for metabolic gene networks under varying environments, we calculated the delta epistasis, which was defined as epistasis differences of the same gene pairs, between each of 16 environment perturbations and the optimum glucose medium that didn't have any environmental perturbations. To avoid including too many gene pairs without any epistasis changes after transitioning from rich medium to harsh environments, we restricted our

analyses to gene pairs with delta epistasis ≥0.001 under at least one of the 16 environmental perturbations. Figure 5.1 provides 4 examples for the first 4 environmental perturbations following the alphabetic order of the 16 conditions. The color bar represents the delta epistasis values under each of the 16 environmental perturbations, in which red and yellow represent increasing epistasis values and blue represents decreasing epistasis values from the rich medium to environmental perturbations.

The results prompted us to speculate that whereas a large proportion of epistasis is rewired dynamically under a variety of environments, there are also epistatic interactions that are conserved under these environmental perturbations. To further quantify our speculations, we divided all epistatic relations into the following five categories: extremely conserved (epistatic relations for the same gene pairs exist in all 16 conditions); high (epistatic relations for the same gene pairs exist in 11-15 conditions); medium (epistatic relations for the same gene pairs exist in 6-10 conditions); low (epistatic relations for the same gene pairs exist in 2-5 conditions); extremely dynamic (epistatic relations for the same gene pairs exist in only 1 condition). We then calculated the ratio of these five types of epistatic relations in each of the 16 environmental perturbations. As shown in Figure 5.2, we found that in each environment, ~40-60% of epistatic interactions were conserved. In addition, different environmental perturbations had very different numbers of condition-specific epistasis. For example, conditions with acetaldehyde and acetate generated much less dynamic epistasis than that of adenosyl methionine and low glucose perturbations.

Figure 5.1. Maps of global epistatic landscapes under environmental perturbations. Different colors represent delta epistasis values as indicated by the color bar to the right. The same colors for the same gene pairs represent the conserved epistatic relations, while different colors indicate that the same gene pairs change epistasis under varying environmental perturbations.

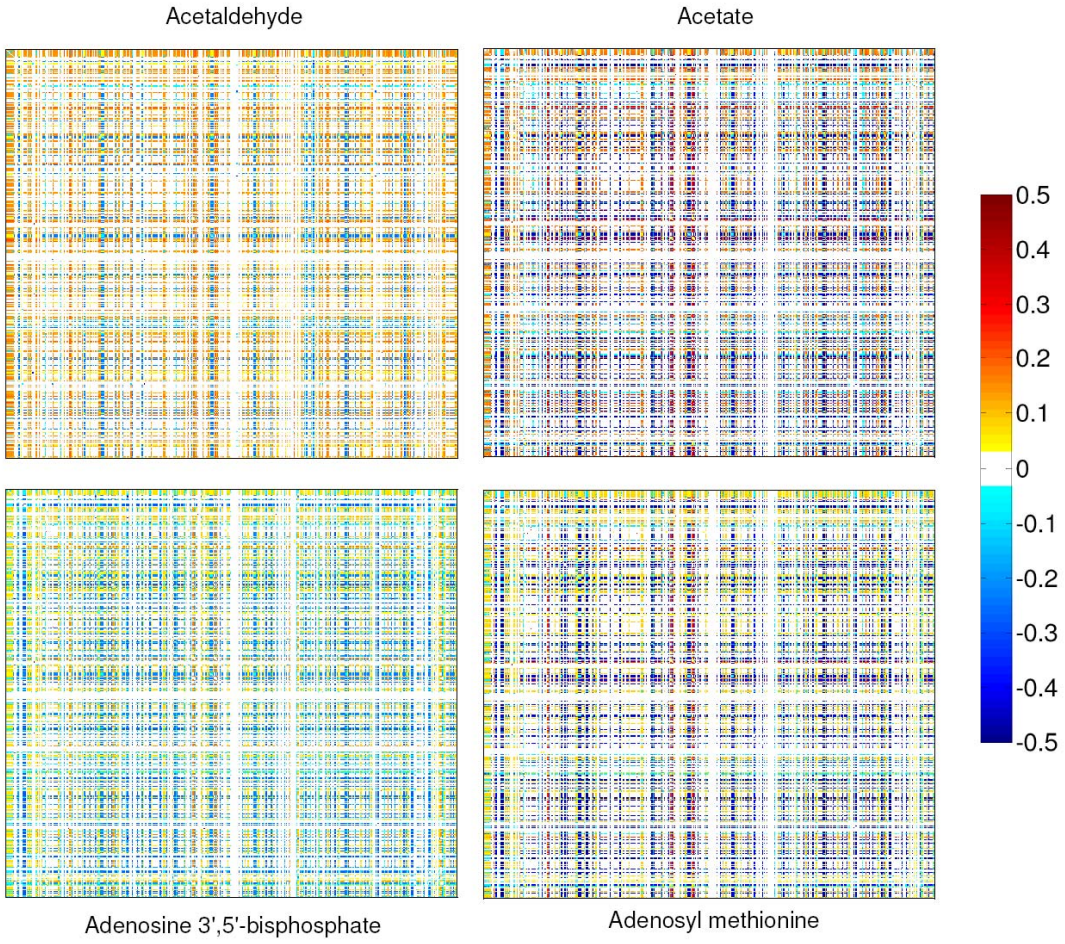
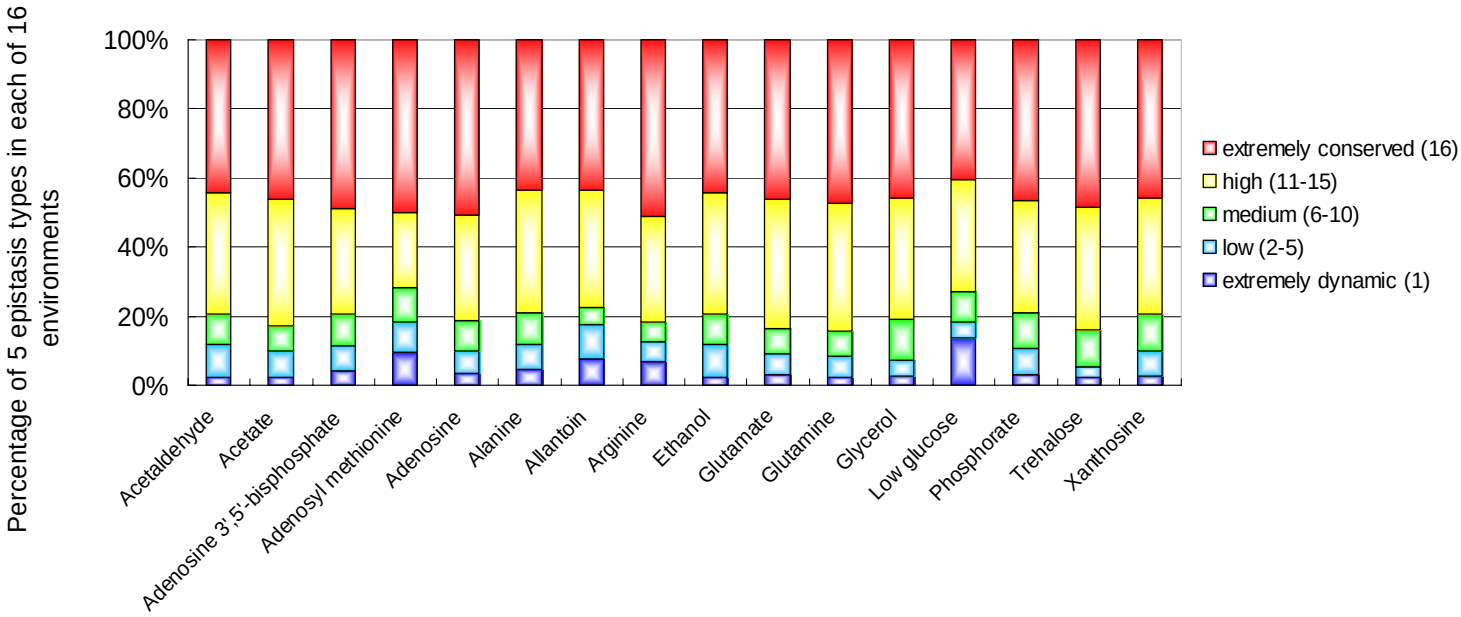

Acetaldehyde
Acetate
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
Adenosine 3',5'-bisphosphate
Adenosyl methionine

Figure 5.2. Fraction of 5 types of epistatic relations in each of the 16 environmental perturbations, as indicated by the color bar to the right. Our results revealed that extremely conserved and extremely dynamic epistasis exist under various environmental conditions.

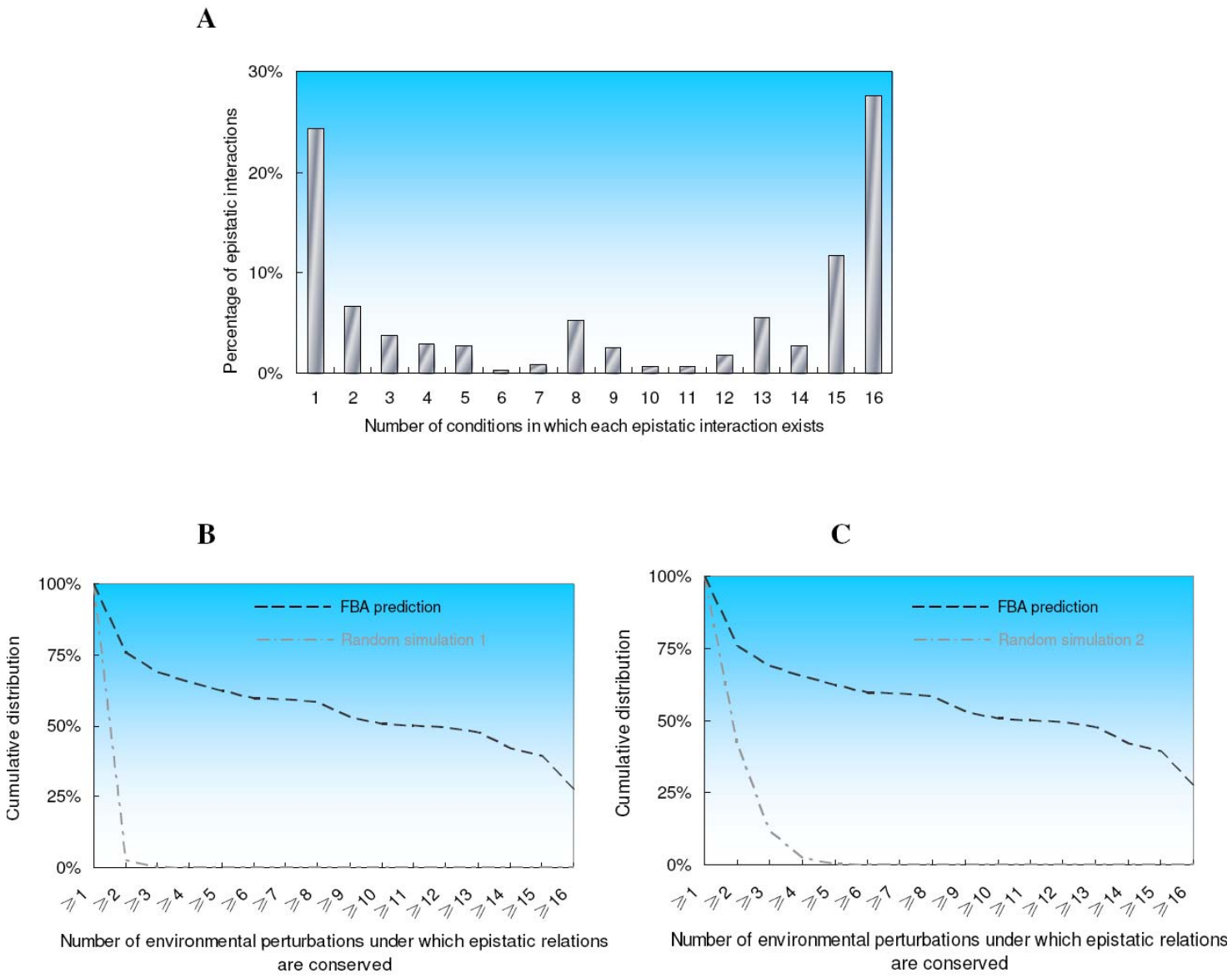

Percentage of 5 epistasis types in each of 16 environments
100%
80%
60%
40%
20%
0%
Acetaldehyde
Acetate
Adenosine 3',5'-bisphosphate
Adenosyl methionine
Adenosine
Alanine
Allantoin
Arginine
Ethanol
Glutamate
Glutamine
Glycerol
Low glucose
Phosphorate
Trehalose
Xanthosine
extremely conserved (16)
high (11-15)
medium (6-10)
low (2-5)
extremely dynamic (1)

5. 3. 2. U-shape distribution of epistasis under environmental perturbations

To understand the global distribution of all epistatic relations, we considered 16 conditions together and calculated the fraction of epistatic interactions existing in 1, 2, 3, ..., 15, 16 conditions, respectively. Interestingly, we found both extremely dynamic and extremely conserved epistasis were most enriched (Figure 5.3): 24% of all epistasis is extremely dynamic, 28% is extremely conserved, and 48% is intermediate (exists in multiple but not all 16 environments). As a result, the U-shape distribution represents a previously unknown pattern that approximately half of epistatic interactions are either extremely dynamic or extremely conserved under a variety of environmental perturbations.

In order to understand whether this pattern could be due to the randomization of the epistatic networks, we compared the distributions based on FBA predictions (Figure 5.3B and 5.3C, black broken lines) with the distributions generated by two distinct randomization strategies (Figure 5.3B and 5.3C, silver broken lines). Firstly, we randomly rewired epistatic relations under each of the 16 conditions, while assuming that the total number of epistatic interactions in each network are constants. As a result, epistatic networks were randomly connected and the number of conditions under which each epistatic interaction exists was calculated. As shown in Figure 5.3B, the distribution of dynamic and conserved epistasis predicted by FBA modeling is very different from that of random simulations: compared to random networks, there is much less extremely dynamic epistasis but more extremely conserved epistasis in FBA predictions.

Figure 5.3. The global distribution of dynamic and conserved epistatic relations under all 16 environments. (**A**) The U-shaped distribution indicates the extremely dynamic and extremely conserved epistatic relations are enriched under environmental perturbations. Note that 28% of epistatic relations are extremely conserved (last bar from left) and 24% are extremely dynamic (first bar from left). (**B**) and (**C**) represent two different strategies for random simulations of epistatic relations under all 16 environments, as indicated in the text. The random simulations were repeated 1,000 times.


A
30%
20%
10%
0%
Percentage of epistatic interactions
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Number of conditions in which each epistatic interaction exists
B
100%
75%
50%
25%
0%
Cumulative distribution
FBA prediction
Random simulation 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Number of environmental perturbations under which epistatic relations are conserved
C
100%
75%
50%
25%
0%
Cumulative distribution
FBA prediction
Random simulation 2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Number of environmental perturbations under which epistatic relations are conserved

In addition, we used another strategy to generate random networks. We randomly rewired epistatic relations while assuming that the total number of epistatic interactions for each gene in each network are constants. As a result, we exchanged interacting partners for any two genes to maintain the connecitivies of each gene and randomly reconnect its partners. We then calculated the number of conditions under which each epistatic interaction exists. The same conclusion, that there is much less extremely dynamic epistasis but more extremely conserved epistasis in FBA predictions, still holds (Figure 5.3C).

5. 3. 3. Co-evolution of genes with epistasis under environmental perturbations

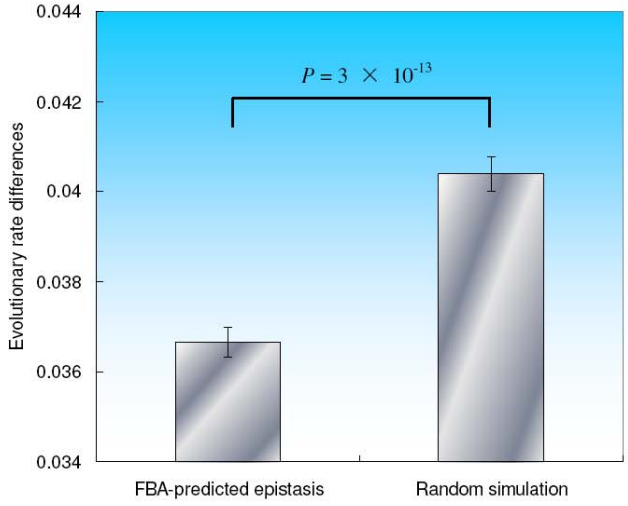
Since above results indicate that epistasis predicted by FBA modeling displays different global distribution patterns from random simulations, we wondered whether the predicted epistasis could have real biological significance. More specifically, we asked whether two genes with predicted epistasis tend to co-evolve since epistasis between genes usually has strong evolutionary consequences (1-9). We calculated the evolutionary rate differences between two interacting genes from FBA modeling (Figure 5.4A). Evolutionary rates (dN/dS) of genes in yeast *S. cerevisiae* were downloaded from Wall et al. (2005) (34) (dN: the nonsynonymous substitutions per synonymous site; dS: the synonymous substitutions per synonymous site). Random simulations with the same number of gene pairs as FBA predictions were conducted to estimate the evolutionary rate differences for any two randomly selected genes. As shown in Figure 5.4A, the gene pairs with FBA-predicted epistatic interactions tend to have a higher chance to co-evolve than random expectation, confirming that our FBA predictions could capture

real biological characteristics of epistatic interactions.

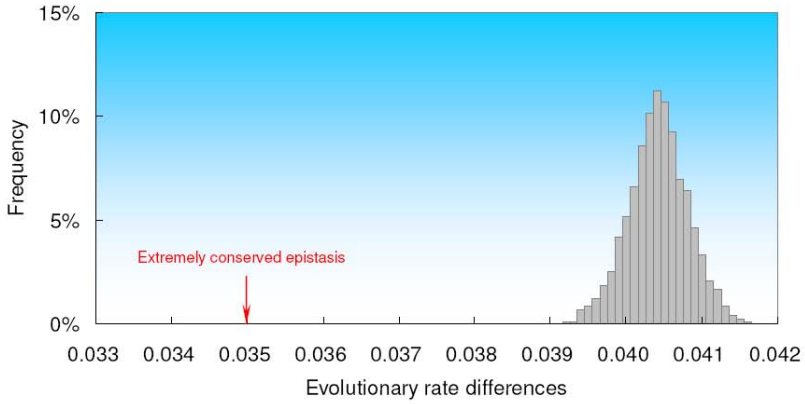
Since Figure 5.3 revealed the unique distribution for extremely conserved and extremely dynamic epistatic interactions, we further compared the evolutionary rate differences between genes with extremely conserved and extremely dynamic epistatic relations. As shown in Figure 5.4B, genes with extremely conserved epistasis tend to co-evolve compared to randomly selected gene pairs ($P < 10^{-4}$), while the difference between genes with extremely dynamic epistasis and random expectation becomes much smaller ($P = 0.06$, Figure 5.4C). The difference between gene pairs with extremely conserved and extremely dynamic epistasis is also significant (t-test, $P = 8\times10^{-6}$). Since extremely conserved epistasis represents epistatic relations that exist under all 16 environmental perturbations, it is more likely to shape the co-evolution relations between two interacting partners than the transient epistasis (e.g. extremely dynamic epistasis) between genes.

Figure 5.4. Co-evolution between gene pairs with FBA-predicted epistasis. (**A**) The random simulations were repeated 10,000 times. The error bars represent standard errors; (**B**) Gene pairs with extremely conserved epistasis tend to co-evolve. Grey bars indicate random simulations with the same number of gene pairs as genes with extremely conserved epistasis. The random simulations were repeated 10,000 times. The red arrow represents the evolutionary rate differences for gene pairs with extremely conserved epistasis, which is significantly different from random simulation (the grey bins) ($P < 10^{-4}$). (**C**) Gene pairs with extremely dynamic epistasis have a weak co-evolving pattern. Grey bars indicate random simulations that were repeated 10,000 times. The blue arrow represents the evolutionary rate differences for gene pairs with extremely dynamic epistasis, which is not significantly different from random simulation ($P = 0.06$).

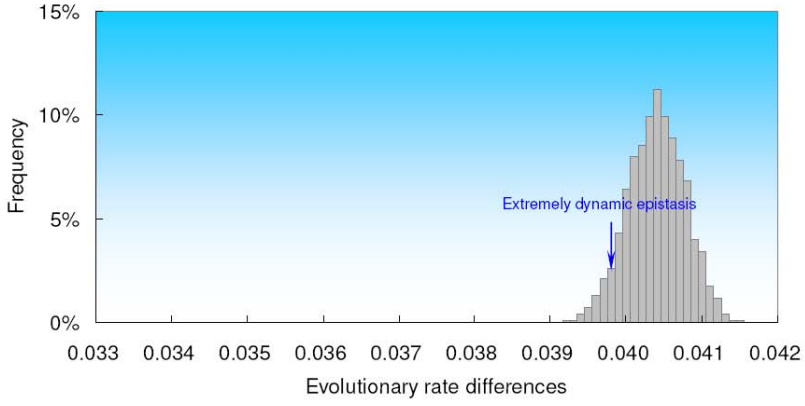
A


0.044
0.042
0.04
0.038
0.036
0.034
$P = 3 \times 10^{-13}$
Evolutionary rate differences
FBA-predicted epistasis
Random simulation

B


15%
10%
5%
0%
Frequency
Extremely conserved epistasis
0.033
0.034
0.035
0.036
0.037
0.038
0.039
0.040
0.041
0.042
Evolutionary rate differences

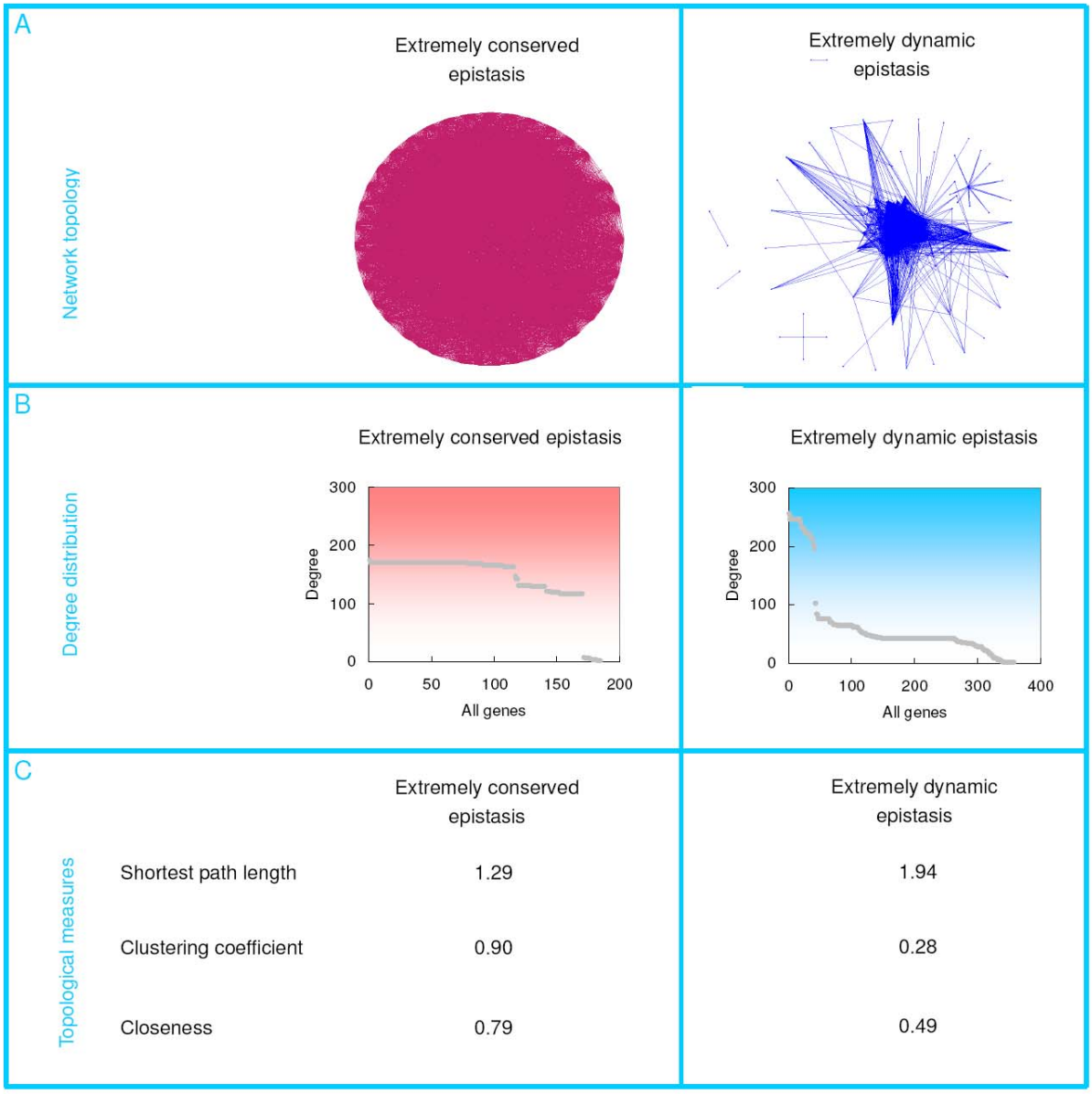
C


15%
10%
5%
0%
Frequency
Extremely dynamic epistasis
0.033
0.034
0.035
0.036
0.037
0.038
0.039
0.040
0.041
0.042
Evolutionary rate differences

5. 3. 4. Network properties of extremely conserved and dynamic epistasis

We further compared extremely conserved and extremely dynamic epistasis and asked whether they had distinct network properties. As shown in Figure 5.5, our results show that extremely conserved epistatic interactions form an exponential network architecture, while extremely dynamic epistatic interactions form a scale-free network topology (35). The exponential network is homogeneous; most nodes have very similar number of links, as shown in the network with extremely conserved epistatic interactions (~120-160 links, Figure 5.5B, left panel). In contrast, the scale-free network is more heterogeneous, where the majority of the nodes have few links but a few nodes have a large number of links, as shown in the network with extremely dynamic epistatic interactions (Figure 5.5B, right panel). In addition, we also calculated three network parameters to compare these two types of epistasis: (1) Shortest path length between two genes reflects the overall network interconnectedness; the smaller the average shortest path length is, the higher chance that genes in this network could interact with the other genes directly; (2) Clustering coefficient is a measurement of the degree to which nodes in a network tend to cluster together; the larger the average clustering coefficient is, the more closely the genes are connected to form modules; (3) Closeness measures the centrality of nodes within a network; nodes that occur on shortest paths with other nodes have higher closeness than those that do not (reviewed in ref. 36). We found that gene pairs with extremely conserved epistasis tend to have smaller shortest path length, larger clustering coefficient and larger closeness than genes with extremely dynamic epistasis (Figure 5.5). These results are consistent with a scenario that genes with extremely conserved epistasis are directly linked to most other genes and form an exponential

Figure 5.5. Network properties of gene pairs with extremely conserved and extremely dynamic epistasis. (**A**) network topologies for extremely conserved and extremely dynamic epistatic interactions respectively; (**B**) degree distribution for gene pairs with extremely conserved and extremely dynamic epistatic interactions respectively; (**C**) three network parameters for two types of epistatic interactions.


A
Network topology
Extremely conserved epistasis
Extremely dynamic epistasis
B
Degree distribution
Extremely conserved epistasis
Extremely dynamic epistasis
Degree
All genes
0
50
100
150
200
300
400
C
Topological measures
Extremely conserved epistasis
Extremely dynamic epistasis
Shortest path length
1.29
1.94
Clustering coefficient
0.90
0.28
Closeness
0.79
0.49

network topology, while genes with extremely dynamic epistasis form a network with a small number of highly connected hub genes and a large number of genes with low connectivities.

5. 4. Discussion

Our study represents the first genome-wide theoretical survey on the dynamics of global epistatic effects under a variety of environmental perturbations. We showed that epistatic profiling at the genome level could be extremely conserved or extremely dynamic under various environmental perturbations. In addition, we found that gene pairs with FBA-predicted epistasis have significantly more similar evolutionary rates, and thus a higher chance to co-evolve than randomly selected gene pairs, confirming that FBA simulation is a useful tool for predicting epistatic interactions. As a result, the observed pattern could represent a previously unknown design principle for the epistatic landscapes under multiple environmental perturbations.

Another interesting finding is that the enrichment of extremely conserved epistasis is unexpected under various strategies of randomization of the epistatic networks, indicating that these extremely conserved epistatic interactions might be responsible for important biological functions. In addition, gene pairs with these extremely conserved epistatic interactions in FBA modeling also have a significantly higher similarity of evolutionary rate than randomly selected gene pairs and genes with extremely dynamic epistasis, which is consistent with a scenario that extremely conserved epistasis plays an important role in long-term evolution. Furthermore, we also found that extremely conserved epistatic interactions forms a distinct network topology from that of extremely dynamic epistasis, indicating that these two types of epistasis play different roles in epistatic interaction networks.

Although we found several novel characteristics regarding the global epistatic landscape under various environmental perturbations, two caveats need to be addressed. First, the FBA modeling used in this study, which was proven to have great predictive power and had been successfully applied to various research problems (e.g. ref. 37-39), only includes metabolic genes in the simulation. Second, even though FBA is one of most comprehensive computational tools for simulating epistatic interactions among genes, there are still many aspects that can be improved to aid in capturing the full set of empirical genetic interactions (40). In the future, we shall integrate rules for transcriptional regulation and physical interactions into this framework to improve over the current FBA methods in predicting epistasis (41).

With these limitations in mind, our observations identified several important features for epistatic interactions under a variety of environmental perturbations, and called on future effort to examine the landscape of epistasis under a wide array of environmental perturbations based on high-throughput experimental platforms. More importantly, the enrichment of conserved and dynamic epistasis would provide a new perspective to understand how biological systems may rewire epistatic interactions under environmental perturbations.

5. 5. Methods

Flux Balance Analysis

Mutations employed in this analysis restricted the flux to be 50% of the wild-type flux found by geometricFBA for all reactions associated with the mutant gene. To find new environments with a specified carbon source or other limiting nutrient that achieves 20% of the high-glucose growth rate, we can solve a linear program for the minimization of the limiting nutrient uptake while requiring the growth rate to be equal to 20% of the rich-glucose growth-rate.

Evolutionary rates and network parameters

Evolutionary rates of genes in *S. cerevisiae* were downloaded from ref. 34. Network parameters such as the shortest path length, clustering coefficient and closeness were calculated using the computer software Pajek, downloaded from: http://vlado.fmf.uni-lj.si/pub/networks/pajek.

References

1. Kondrashov AS (1982) Selection against harmful mutations in large sexual and asexual populations. *Genet Res.* 40:325-332.

2. Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL (2006) Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* 440:87-90.

3. Otto SP (2007) Unraveling the evolutionary advantage of sex. *Genet Res.* 89:447-449.

4. Presgraves DC (2007) Speciation genetics: epistasis, conflict and the origin of species. *Curr Biol.* 17:R125-R127.

5. Hansen TF, Wagner GP (2001) Epistasis and the mutation load: a measurement theoretical approach. *Genetics* 158:477-485.

6. Kondrashov AS, Crow JF (1991) Haploidy or diploidy: which is better? *Nature* 351:314-315.

7. Musso G, et al. (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18:1092-1099.

8. Perez-Figueroa A, Caballero A, Garcia-Dorado A, Lopez-Fanjul C (2009) The action of purifying selection, mutation and drift on fitness epistatic systems. *Genetics* 1831:299-313.

9. Sanjuan R, Nebot MR (2008) A network model for the correlation between epistasis and genomic complexity. *PLoS One* 37:e2663.

10. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I (2009) Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet.* 5:e1000578.

11. Phillips PC (2008) Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 9:855-867.

12. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327:425-431.

13. Tong AH, et al. (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808-813.

14. Pan X, et al. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol Cell* 16:487-496.

15. Pan X, et al. (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 124:1069-1081.

16. Measday V, Hieter P (2002) Synthetic dosage lethality. *Methods Enzymol* 350:316-326.

17. Measday V, et al. (2005) Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc Natl Acad Sci USA* 102:13956-13961.

18. Sopko R, et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21:319-330.

19. Collins SR, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446:806-810.

20. Kornmann B, et al. (2009) An ER-mitochondria tethering complex revealed by a synthetic biology screen. *Science* 325:477-481.

21. Fiedler D, et al. (2009) Functional Organization of the *S. cerevisiae* Phosphorylation Network. *Cell* 136:952-963.

22. Bandyopadhyay S, et al. (2010) Rewiring of Genetic Networks in Response to DNA Damage. *Science* 330:1385-1389.

23. Becker SA, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2:727-738.

24. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28:245-248.

25. Smallbone K, Simeonidis E (2009) Flux balance analysis: a geometric perspective. *J Theor Biol* 258:311-315.

26. Edwards JS, Ibarra RU, Palsson BØ (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19:125-130.

27. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 102:7695-7700.

28. AbuOun M, et al. (2009) Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. *J Biol*

Chem 284:29480-29488.

29. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* 33:164-190.

30. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* 26:659-667.

31. Fong SS, et al. (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 91:643-648.

32. Fong SS, Joyce AR, Palsson BØ (2005) Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res.* 15:1365-1372.

33. Mo ML, Palsson BØ, Herrgard MJ (2009) Connecting extracellular metabolomic profiles to intracellular metabolic states in yeast. *BMC Syst Biol* 3:37.

34. Wall DP, et al (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* 102:5483-5488.

35. Albert R, Jeong H, Barabasi A-L (2000) Error and attack tolerance of complex networks. *Nature* 406:378-382.

36. Barabási A-L, Oltvai Z (2004) Network Biology: Understanding the Cells's Functional Organization. *Nat Rev Genet* 5:101-113.

37. Papp B, Pal C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429:661-664.

38. Segrè D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* 37:77-83.

39. He X, Qian W, Wang Z, Li Y, Zhang J (2010) Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat Genet* 42:272-276.

40. Covert MW, Schilling CH, Palsson BØ (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* 213:73-88.

41. Szappanos B, et al. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* 43:656-662.

42. Schellenberger J, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0 *Nature Protoc* 6:1290–1307